Gender Bias in Artificial Intelligence

A Systematic Review of the Literature

Rosileine Mendonça de Lima
Paulista University, São Paulo, Brazil

Barbara Pisker
Josip Juraj Strossmayer University of Osijek, Croatia

Victor Silva Corrêa
Paulista University, São Paulo, Brazil

Abstract: This study presents a Systematic Literature Review (SLR) of Gender Bias in Artificial Intelligence (AI). The research was conducted using two techniques: a domain-based approach to SLR process providing a bibliometric sample description and in-depth examination of the thematic categories arising from inductive categorization, extracted from reading and interpretation of the final 35 sample articles analyzed. In answering three key research questions on the types, causes, and overcoming (mitigating) strategies of gender bias in artificial intelligence, three thematic treemaps were constructed, enabling systematic overview as an essential contribution to the literature. The main types of gender bias found in AI are categorized as societal, technical, and individual. Societal and socio-technical aspects stand out as the leading causes of bias, while debiasing, dataset design and gender sensitivity were the most frequent among the main strategies for overcoming bias. The study also proposes theoretical, practical and managerial capacity building and policy implications that aim to influence broad socio-technical challenges and refer to changes necessary, aiming to create bias-free artificial intelligence.

Keywords: Bias, Gender, Artificial Intelligence, Systematic Literature Review

Introduction

If we think in terms of creation, drawing a parallel with the thesis in the first book of Moses called Genesis, just as God created man in his image, nowadays we live in a tech society witnessing profound developments in AI, in the role of a creator as humankind. Besides acknowledging the undeniable fact that we live in thrilling times of acceleration, as we

http://doi.org/10.18080/jtde.v11n2.690

approach the point of singularity, we are obliged to reassure ourselves that great power also comes with great responsibility.

Being aware of the social responsibility we bear in creating and using AI, we need to question, test, and debate its potentially embedded biases starting from within our human imperfection, especially being aware of the biases coded in a creator's mindset. For this systematic literature review (SLR) research, we focused on gender bias in AI, aiming to answer three main research questions (RQ):

RQ1) What are the main types of gender bias in AI?

RQ2) What are the leading causes of these AI gender biases?

RQ3) What are the main strategies for overcoming (mitigating) gender biases in AI?

Early work on the topic of the intersection between gender and AI followed an overall, broader socio-generic pattern predominantly in tech sciences literature, starting with Licklider and Taylor (1968) discussing the potential impact of computer technology on society, and raised concerns regarding the reinforcement of existing social biases and discrimination. Further, Deacon and Brooks (1988) argued that the biases and limitations of human designers and programmers could be reflected in artificial intelligence systems and have negative consequences for their users. Finally, Breazeal and Brooks (1997) examined the impact of gender biases in artificial intelligence and robotics research and development and called for more diverse and inclusive approaches to these fields.

In feminist theory, we find early works by Haraway (1987, 1991), Turkle (2005), and Oldenziel (1992) focused closely and specifically on gendered aspects of technology, gender-biased technology embedment, and their societal implications. Haraway's work laid the foundation for feminist discussions on technology, including the gendered implications of biased AI. Gender-focused AI literature development continues in social sciences in the works of Wajcman (2004), Crawford (2013, 2021), and Noble (2018), gaining full empirical materialization and entering the vivid scientific debate in the last quinquennial.

As AI is becoming an omnichanger in contemporary societies, with the interdisciplinary scientific research literature on gender-biased AI continuing to grow exponentially, it is scientifically justified to systematize the literature contributing towards a comprehensive review in terms of critical causes, types, and mitigating strategies on gender biases in AI presented in this research paper. Previous systematic literature reviews on gender-biased AI have just begun to develop in the field, adding significant contributions. Kordzadeh and Ghasemaghaei (2022) deliver a review, synthesis and future research directions, Reyero Lobo *et al.* (2022) show the applicability of semantics to address bias in AI, while Fyrvald (2019) suggests solutions for mitigating algorithmic bias in AI systems based on qualitative research.

http://doi.org/10.18080/jtde.v11n2.690

Nadeem *et al.* (2022) conceptualized gender bias in AI-based decision-making systems, proposing mitigating strategies for biased effects. There are also relevant contributions from Wellner and Rothman (2020) regarding the idea of feminist AI and a relevant overview of the state of gender equality in and by AI from Patón-Romero *et al.* (2022).

Although recent years have shown a growing trend in the body of literature and research on gender bias in artificial intelligence, especially in the last decade, the field lacks systematization, broader research network interest and rootedness in the social sciences field. The paper aims to contribute towards a deeper understanding of the current state of knowledge on this topic, categorizing types, causes and overcoming strategies for gender bias in AI. While providing insights into the most effective strategies for addressing gender bias in AI, the paper highlights the need for further research in this area. Additionally, it provides a valuable resource for policymakers, practitioners, and other stakeholders, recommending best practices for overcoming (mitigating) gender bias in AI. The research also has the potential to significantly advance understanding of this critical issue in the field of social sciences and set a path for future research.

Method

The authors used a systematic literature review (SLR) as a research method that relies primarily on content analysis for inductive data extraction (Kraus *et al.*, 2020). As proposed here, SLR can be employed for various purposes, including collating, synthesizing, and mapping literature in the field. Indeed, the SLR proved to be adequate for answering these three research questions. In addition to the content analysis in response to the questions, this study also performed a bibliometric analysis, describing the articles according to their publication incidence by year, author, and country (Paul & Criado, 2020). In addition to content analysis in response to the questions, this study also performed a bibliometric description of the final sample, describing the articles according to their publication incidence by year, author, and country (Corrêa *et al.*, 2022b; Paul & Criado, 2020).

Search strategy

The authors heeded the three stages proposed by Tranfield *et al.* (2003): planning, conducting, and disseminating. In the planning stage, the authors created a research protocol after identifying theoretical and empirical gaps that suggested the proposal's relevance (Table 1) (Machado *et al.*, 2020; Tranfield *et al.*, 2003). Next, the authors defined the criteria that would include and exclude articles and the quality aspects of the papers that should be considered when selecting the final sample. Based on these criteria, the authors filtered and set the final research sample. In the third stage, disclosure, the authors conducted an

http://doi.org/10.18080/jtde.v11n2.690

explanatory and in-depth examination of the thematic categories arising from inductive categorization, that is, extracted from the reading and interpretation of the articles in light of the proposed questions (Machado *et al.*, 2020; Tranfield *et al.*, 2003). In a systematic review of the literature on female entrepreneurship in emerging and developing contexts, Corrêa *et al.* (2022a, p. 306) defend the relevance of inductive and exhaustive categorization, as they allow for eventual discoveries that may not express or reframe "a developed theoretical research stream". Indeed, several authors have shed light on the relevance of systematic reviews based on inductive thematic categorizations, allowing insights and discoveries from the underlying literature that are not plausible in deductive models (Conz & Magnani, 2020; Hägg & Gabrielsson, 2020; Mahmud *et al.*, 2022; Santos & Neumeyer, 2021). Following Conz & Magnani (2020, p. 402), we categorize articles employing "inductive qualitative content analysis, adopting the so-called 'conventional approach' to the coding process, which is generally used in studies whose aim is to describe a phenomenon, when existing theory or research literature is limited". Still following these authors, "we performed the content analysis individually and then discussed the results together, confronting emerging categories and subcategories of descriptions" (Conz & Magnani, 2020, p. 402). Aiming to allow readers to replicate the results and the primary and secondary categorizations identified here, we make available the truth matrix, containing all primary and secondary categories obtained from the inductive analysis of the evidence, enhancing the validity and reliability of the study. The truth matrix is available as a permanent link through DOI https://doi.org/10.6084/m9.figshare.22811450.v1.

Table 1. Research protocol

Research protocol	Detailed description
Research various databases	Scopus Database and Web of Science
Publication Type	Peer-review journals
Language	English
Date Range	2012-2022.
Search fields	Title, abstract, and keywords
Search terms (*Scopus*)	(TITLE-ABS-KEY ("Artificial intellige*") OR TITLE-ABS-KEY ("machine learning") OR TITLE-ABS-KEY ("natural language processing") OR TITLE-ABS-KEY ("neural networks") OR TITLE-ABS-KEY (Robotic*) AND TITLE-ABS-KEY ("gender bias") OR TITLE-ABS-KEY ("gender disparity") OR TITLE-ABS-KEY ("gender imbalance") OR TITLE-ABS-KEY ("gender inequality")) AND (LIMIT-TO (SUBJAREA, "SOCI") OR LIMIT-TO (SUBJAREA, "ARTS") OR LIMIT-TO (SUBJAREA, "PSYC") OR LIMIT-TO (SUBJAREA, "MULT") OR LIMIT-TO (

Research protocol	Detailed description
	SUBJAREA, "DECI") OR LIMIT-TO (SUBJAREA, "BUSI") OR LIMIT-TO (SUBJAREA, "ECON")) AND (LIMIT-TO (DOCTYPE, "ar")) AND (LIMIT-TO (LANGUAGE, "English"))
Search terms (*Web of Science*)	(Topic ("Artificial intellige*") OR Topic ("machine learning") OR Topic ("natural language processing") OR Topic ("neural networks") OR Topic (“Robotic*”) AND Topic ("gender bias") OR Topic ("gender disparity") OR Topic ("gender imbalance") OR Topic ("gender inequality")) AND (LIMIT-TO (SUBJAREA , "Multidisciplinary Sciences") OR LIMIT-TO (SUBJAREA , "Ethics") OR LIMIT-TO (SUBJAREA , "Communication") OR LIMIT-TO (SUBJAREA , "International Relations") OR LIMIT-TO (SUBJAREA , "Language Linguistics") OR LIMIT-TO (SUBJAREA , "Linguistics") OR LIMIT-TO (SUBJAREA , "Philosophy") OR LIMIT-TO (SUBJAREA , "Psychology Social") OR LIMIT-TO (SUBJAREA , "Political Science") OR LIMIT-TO (SUBJAREA , "Sociology") OR LIMIT-TO (SUBJAREA , "Humanities Multidisciplinary") OR LIMIT-TO (SUBJAREA , “Women S Studies”) OR LIMIT-TO (SUBJAREA , "Development Studies") OR LIMIT-TO (SUBJAREA , "Environmental Studies") OR LIMIT-TO (SUBJAREA , "Regional Urban Planning") OR LIMIT-TO (SUBJAREA , "Urban Studies") OR LIMIT-TO (SUBJAREA , "History Philosophy Of Science") OR LIMIT-TO (SUBJAREA , "Education Scientific Disciplines")) AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (LANGUAGE , "English"))
Inclusion criteria	Article. Articles published in English.
Exclusion criteria	Grey literature (conference-published papers, non-peer-reviewed works); Works published in languages other than English.

Selection criteria

The following criteria guided the final selection of the articles: First, the articles should be listed in Scopus or Web of Science (WoS) “because their coverage and selective approach produce a curated collection of documents” (Machado *et al.*, 2020). The articles were searched in October 2022, considering the criteria presented in Table 1. Following Antony *et al.* (2020) and Corrêa *et al.* (2022a, 2022b), we included articles published within the last ten years of the search base date, considering only articles published in English.

Search process

After a previous search for articles related to gender bias and artificial intelligence, equivalent terms and/or most used synonyms were identified: “artificial intellige*”; “machine learning”; “natural language processing”; “neural networks”; “robotic*”; “gender bias”; “gender disparity”; “gender imbalance” and “gender inequality”. To achieve broader coverage, the search combined existing terms in the abstract, title or keywords with the Boolean operator

http://doi.org/10.18080/jtde.v11n2.690

"OR". The search yielded 121 articles (69 from Scopus, 52 from WoS). Of these, 58 were available in both databases. The authors then used an Excel spreadsheet to eliminate repetitions, leaving 63 articles. We excluded 16 articles unrelated to the social and applied sciences or the purpose of the research, leaving 47 articles. These 47 articles progressed to the next stage, where three authors performed independent readings. The goal was to allow each author to evaluate the articles that should advance to the next step.

Inclusion and exclusion criteria

Initially, the authors selected only articles published in peer-reviewed, open-access journals and written in English between 2012 and 2022. In addition, we excluded grey literature. We then evaluated the quality of the articles and selected 35 for the final sample. Figure 1 illustrates the SLR process.

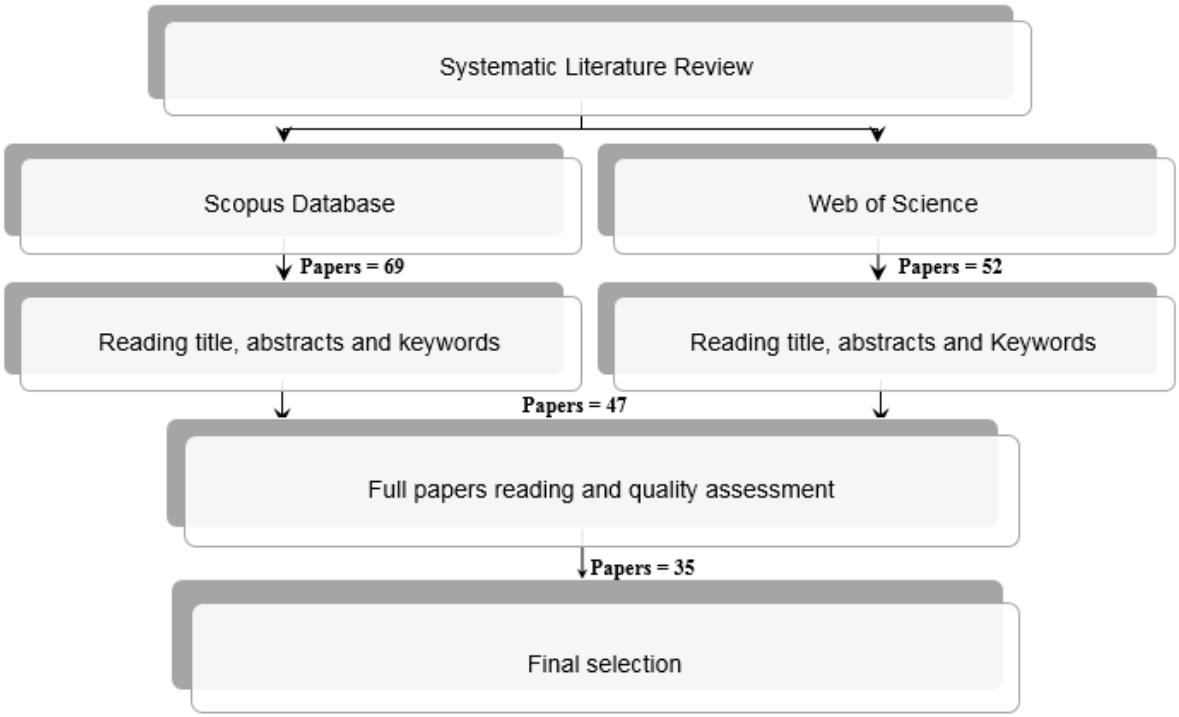

Figure 1. SLR process

Bibliometric Description of the Articles

Papers published by year

Figure 2 displays the publication of the articles per year. It is possible to see how the gender bias in AI themes has grown, with particular emphasis on the last three years (2020–2022). Some aspects stand out. Although the search period initially incorporated 2012 and 2013, no articles were found for either year. Considering the search terms and inclusion and exclusion criteria, the first study was published in 2014, with only one article. If we consider the 63

http://doi.org/10.18080/jtde.v11n2.690

articles in the final sample, 51 (more than 80 %) were published in the last three years (2020–2022), demonstrating a notable growth of interest in the subject by researchers in the area.

It should be noted that the database search was conducted in October 2022; therefore, it only considered some articles on the subject published in 2022. Even considering only eight months of 2022, 22 articles were published on gender bias in AI, which was 38% higher than the number obtained a year earlier. The increase in the number of works expresses the growing relevance of the theme in empirical and academic contexts. Indeed, Shrestha and Das (2022, p. 1) have stressed how "algorithmic fairness has been a topic of interest in the academy for the past decade", including and not restricted to reflections on gender bias. In this context, Asr *et al.* (2021, p. 1) have emphasized how "Women's voices are disproportionately underrepresented" and how this underrepresentation reaches different areas of society. The increase in the number of studies that seek to understand such gaps is a growing concern in understanding this phenomenon. According to Shrestha and Das (2022, p. 1), "although the fairness discussion within the realm of ML [Machine Learning] and AI [Artificial Intelligence] is a recent development, discrimination has roots within human society". Only recently have studies on AI and gender bias begun to shed light on gender bias. According to the authors, "gender bias is most harmful when it is not as readily noticeable" (Shrestha & Das, 2022, p. 2).

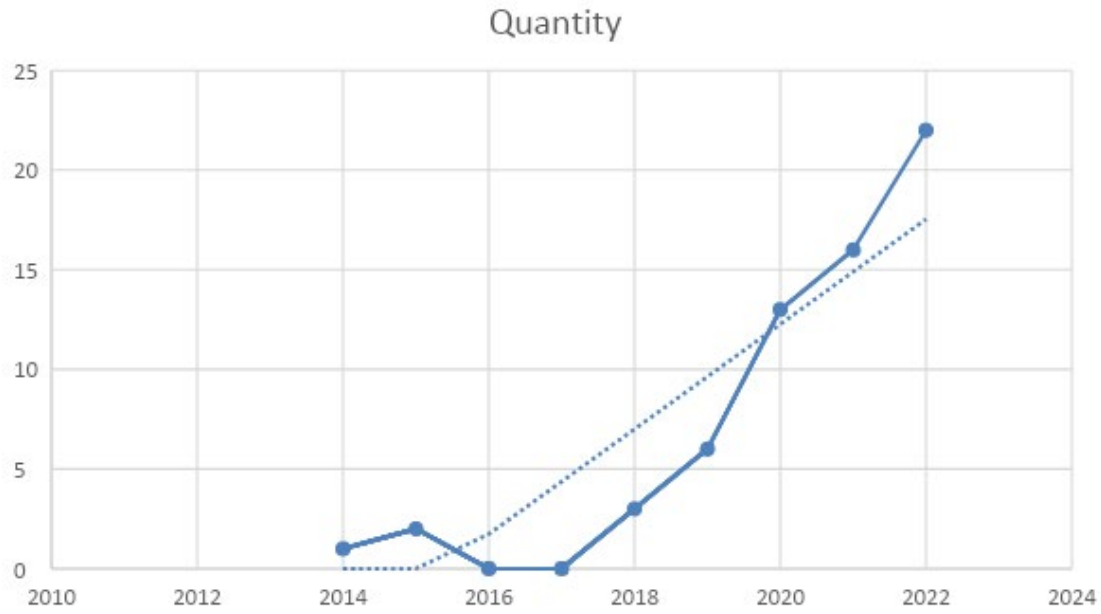

Figure 2. Publication of articles per year

Papers published by country

Table 2 presents the publication of the articles by the country in which the authors conducted the empirical research. The countries were counted individually when a survey was conducted in more than one country. The data indicated that the surveys were mainly conducted in the United States, with five studies. Next, we examine four emerging and developing economies:

http://doi.org/10.18080/jtde.v11n2.690

Chile (2 papers), India (2 papers), Argentina (1 paper), and Bangladesh (1 paper). Considerable attention has been drawn to the fact that out of the 63 articles in the final sample, for 16 of them, it was not possible to identify the place(s) where the research was carried out, suggesting the need for the authors to be more didactic and transparent in the methodological aspects of their studies.

Table 2. Publication of the articles by country

Country	Number
USA	5
Chile	2
India	2
Argentina	1
Bangladesh	1
Canada	1
Bolivia	1
Colombia	1
El Salvador	1
Spain	1
France	1
England	1
Ireland	1
Mexico	1
Nigeria	1
Kenya	1
Switzerland	1
Several	11
Undefined	16

Paper published by author

A total of 228 authors published the papers of the final sample. Author repetition was low among the papers published. Only James Zou published two papers on gender bias in AI. All the other 227 authors published only one article from the final sample, demonstrating prominent fragmentation in the theme study.

http://doi.org/10.18080/jtde.v11n2.690

Discussion and Implications

Regarding RQ1 (What are the main types of gender bias in AI?), Figure 3 presents a table of treemaps for the main types of gender bias in studies related to artificial intelligence. The 34 articles generated 137 records, subdivided into five primary categories: social and technical bias, individual bias, emerging bias, and linguistic bias. Although the inductive categorization generated five categorical levels, Figure 3 illustrates two main levels: primary and secondary. From this figure, it is possible to observe the main categories that came from the inductive analysis of the articles. For example, in addition to the five primary categories, this study identified several secondary subcategories, broadening reflections related to the theme. Figure 3 expands the field's understanding by projecting light onto new reflections that are still little explored by classical or more contemporary authors (Wajcman, 2004; Crawford, 2013, 2021; Noble, 2018; Kordzadeh & Ghasemaghaei, 2022). For example, although more recent studies such as those by Wajcman (2004), Crawford (2013, 2021) and Noble (2018), have advanced the discussion on gender biases and AI, they do not advance or propose related reflections.

Studies on gender bias in the context of artificial intelligence mainly comprised societal bias (88 % [n = 30] of the 34 studies analyzed), generating 115 records associated with this category. Among the societal biases, ten articles emphasized the biases in social structures that can be incorporated into artificial intelligence systems, compromising the results presented. (Vlasceanu & Amodio, 2022; Asr *et al.*, 2021; Bhardwaj *et al.*, 2021; Schopmans & Cupac, 2021; Schwemmer *et al.*, 2020; Fossa & Sucameli, 2022; Petreski & Hashim, 2022; Schwemmer *et al.*, 2020; Tomalin *et al.*, 2021).

The articles also highlighted other biases. For example, six studies sought to understand how human bias influences data generated by AI systems (Kuppler, 2022; Jones *et al.*, 2020; Tomalin *et al.*, 2021; Petreski & Hashim, 2022; Schwemmer *et al.*, 2020; Tannenbaum *et al.* , 2019). In addition, it highlights that pre-existing bias (Savoldi *et al.*, 2021; Draude *et al.*, 2020; Huluba *et al.*, 2018), racial bias (Chen *et al.*, 2022; Draude *et al.*, 2020; Martínez *et al.*, 2020; Scheuerman *et al.*, 2019; Schwemmer *et al.*, 2020; Waelen & Wieczorek, 2022), structural bias (Tubaro *et al.*, 2022; Draude *et al.*, 2020; Martínez *et al.*, 2020; Schopmans & Tupac, 2021; Schwemmer *et al.*, 2020; Tomalin *et al.*, 2021) and male pattern bias (Huluba *et al.*, 2018; Jones *et al.*, 2020; Vlasceanu & Amodio, 2022; Tomalin *et al.*, 2021; Petreski & Hashim, 2022) stand out among the investigated biases.

The second group incorporated those related to technical bias that could influence the responses of AI systems (59% [n=20]). The main preferences in this category include algorithmic (Savoldi *et al.*, 2021; Vlasceanu & Amodio, 2022; Jones *et al.*, 2020; Draude *et al.*, 2020; Thelwall, 2018; Schwemmer *et al.*, 2020; Waelen & Wieczorek, 2022; Tannenbaum *et*

http://doi.org/10.18080/jtde.v11n2.690

al., 2019) and selection biases (Huluba *et al.*, 2018; Jones *et al.*, 2020; Witherspoon *et al.*, 2016).

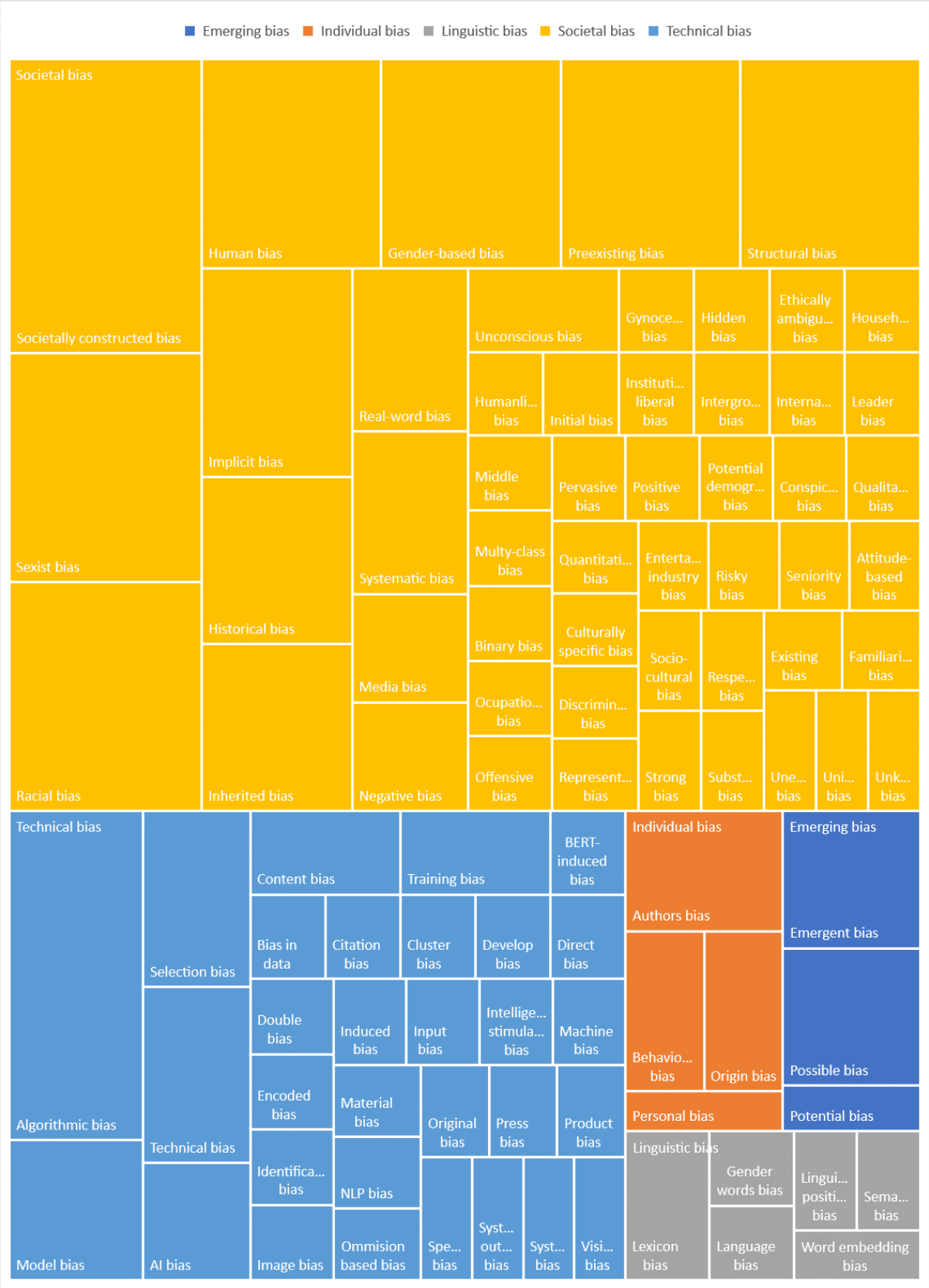

Figure 3. Treemap for the main types of gender biases in studies related to artificial intelligence

http://doi.org/10.18080/jtde.v11n2.690

The third group was related to individual bias (n=8). Among them, factors such as pre-judgment of personal origin (Pair *et al.*, 2019; Kurpicz-Briki & Leoni, 2021; Asr *et al.*, 2021; Bhardwaj *et al.*, 2021; Schopmans & Cupac, 2021), prejudice from researchers and authors (Bardhan *et al.*, 2019; Jones *et al.*, 2020), and behavioural bias (Fossa & Sucameli, 2022) stand out. Finally, the fourth and fifth groups comprised studies that investigated emergent bias (n=6), followed by a linguistic bias (n=5). Studies on emerging biases are mainly related to the potential prejudice generated by the misuse of AI Systems (Asr *et al.*, 2021; *Draude et al.*, 2020; Martínez *et al.*, 2020; Tannenbaum *et al.*, 2019). On the other hand, linguistic bias studies have focused on word bias's impact (Pair *et al.*, 2021; Kurpicz-Briki & Leoni, 2021; Martínez *et al.*, 2020) and language bias (DeFranza *et al.*, 2020; Orgeira-Crespo *et al.*, 2021).

Regarding RQ2 (What are the leading causes of these AI gender biases?), Figure 4 presents a table of treemaps with the leading causes of these biases in the studies related to artificial intelligence. It is important to emphasize that, like the categorization in response to RQ1, the categorization of RQ2 was also inductive and dynamic. Forty-seven articles were read, and only 14 presented the causes; 33 articles described the types of existing biases, but the causes of these occurrences needed to be described. The 14 articles analyzed generated 39 records grouped into six categories: Systems, Prejudice, Culture, Inequality, Relationship, and Interaction. In addition, we classified the articles into more than one category that addressed the different causes of these biases.

Among the leading causes of the gender biases in artificial intelligence, the term “systems” is found most often (in 64% [n=9] of the 14 studies analysed) (Savoldi *et al.*, 2021; Vargas-Solar, 2022; Pair *et al.*, 2021; Chen *et al.*, 2022; Dwork & Minow, 2022; Jones *et al.*, 2020; Schopmans & Cupac, 2021; Scheuerman *et al.*, 2019; Thelwall, 2018; Fossa & Sucameli, 2022). Articles present contexts in which systems may have been developed incorporating constraints and technical decisions (Savoldi *et al.*, 2021; Chen *et al.*, 2022; Fossa & Sucameli, 2022), extraction (Chen *et al.*, 2022; Dwork & Minow, 2022), data patterns (Vargas-Solar, 2022; Pair *et al.*, 2021; Dwork & Minow, 2022; Jones *et al.*, 2020; Schopmans & Cupac, 2021; Scheuerman *et al.*, 2019; Thelwall, 2018), resulting in incomplete or defective AI systems (Vargas-Solar, 2022; Dwork & Minow, 2022; Thelwall, 2018) due to social prejudices and stereotypes transferred to algorithms by programmers (Pair *et al.*, 2021; Thelwall, 2018).

Prejudice was the object of attention in seven of the 14 articles analyzed, generating 13 records. Savoldi *et al.* (2021) and Huluba *et al.* (2018), Vlasceanu and Amodio (2022) and Pair *et al.* (2021), Petreski and Hashim (2022) and Schopmans and Cupac (2021) associated the causes of gender bias with social pressure for a standard that is considered acceptable, which may have influenced the ways AI systems were technically designed. Another striking reason found

http://doi.org/10.18080/jtde.v11n2.690

is the prejudices and stereotypes of AI system programmers (Savoldi *et al.*, 2021; Huluba *et al.*, 2018; Jones *et al.*, 2020; Petreski & Hashim, 2022; Schopmans & Cupac, 2021).

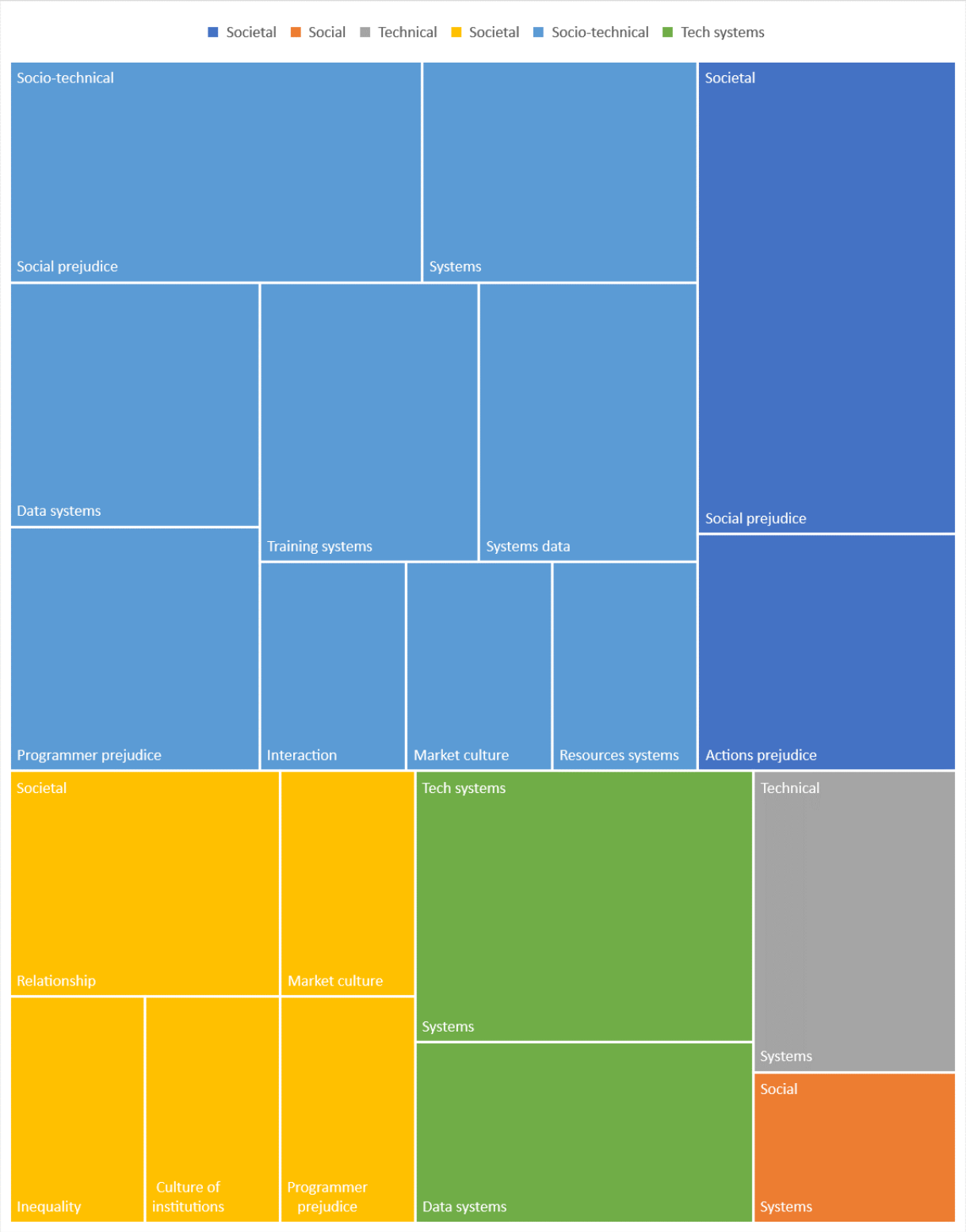

Figure 4. Treemap with the leading causes of gender biases in the studies on artificial intelligence

The third category of cause of bias is linked to the influence of the culture that permeates the internal and external environments in which the system is developed (n=2), such as referring

http://doi.org/10.18080/jtde.v11n2.690

to the structure and composition of the formal labour market (Tubaro *et al.*, 2022), industry-related motives and skills in the sector, and initial institutional choices (Huluba *et al.*, 2018).

Other observed causes were related to inequality in the representation of women in technology professions (n=1), who could be professional programmers of these systems or involved in projects to make them more realistic (Tubaro *et al.*, 2022). Finally, the sixth category (n=1) points to the interaction between the user and the system as a cause of the emergence of biases in AI systems (Savoldi *et al.*, 2021).

Regarding *RQ3* on the main strategies for overcoming (mitigating) gender biases in AI, Figure 5 presents a table of treemaps with the leading strategies for overcoming these biases in the studies related to artificial intelligence. Out of 47 observed articles, only 14 debated the ways to overcome gender bias in artificial intelligence, resulting in 58 overcoming strategies, with 49 different records. Those 49 different extracted records were further organized into 12 categories for overcoming gender bias in artificial intelligence: Debiasing, Dataset design, Gender sensitivity, Inclusiveness, Transparency, Fairness, Sociotechnical entanglements, Word embedding, Monitoring, Regulation, Optimization and Certification.

Among the categories found to overcome gender bias in artificial intelligence, debiasing is most frequent, with 35.71% (n=5) of papers analyzed (Savoldi *et al.*, 2021; Vlasceanu & Amodio, 2022; Kurpicz-Briki & Leoni, 2020; Tomalin *et al.*, 2021). Savoldi *et al.* (2021) presented model debiasing patterns (with gender tagging, adding context, debiased word embeddings and balanced fine tuning) and debiasing through external components (black-box injection, lattice rescoring and gender reinflection). Also, different types of debiasing are found: gender, cultural, dataset, external, language and model debiasing (Bhardwaj *et al.*, 2021; Tomalin *et al.*, 2021). Vlasceanu & Amodio (2022) find bias de-propagation to break the cycle of bias propagation between society and AI, while Kurpicz-Briki & Leoni (2020) find solutions in debiasing through word embeddings.

Dataset design and Gender sensitivity categories follow, both with 28.57% presence share (n=4) encompassing over half of all overcoming solutions for gender bias in AI avoidance. As seen in Vargas-Solar's (2022) study, the dataset design should be completed by inserting missing women's history datasets, while Draude *et al.* (2020) find importance in the balance of dataset nutrition, accountability, context inclusion, fairness, justice, gender stereotypes removal, explainable AI, and dataset diversification. Savoldi *et al.* (2021) highlight domain adaptation, upsampling, downsampling and counterfactual augmentation as pathways for remodelling dataset design in overcoming gender bias in AI. Gender sensitivity is seen as mainstream in Bardhan *et al.* (2019), while gender gap-tracker, context-sensitive gender

http://doi.org/10.18080/jtde.v11n2.690

inference and gender set self-identification are a cornerstone of balancing gender bias in AI (Asr *et al.*, 2021; Das & Paik, 2021; Scheuerman *et al.*, 2019).

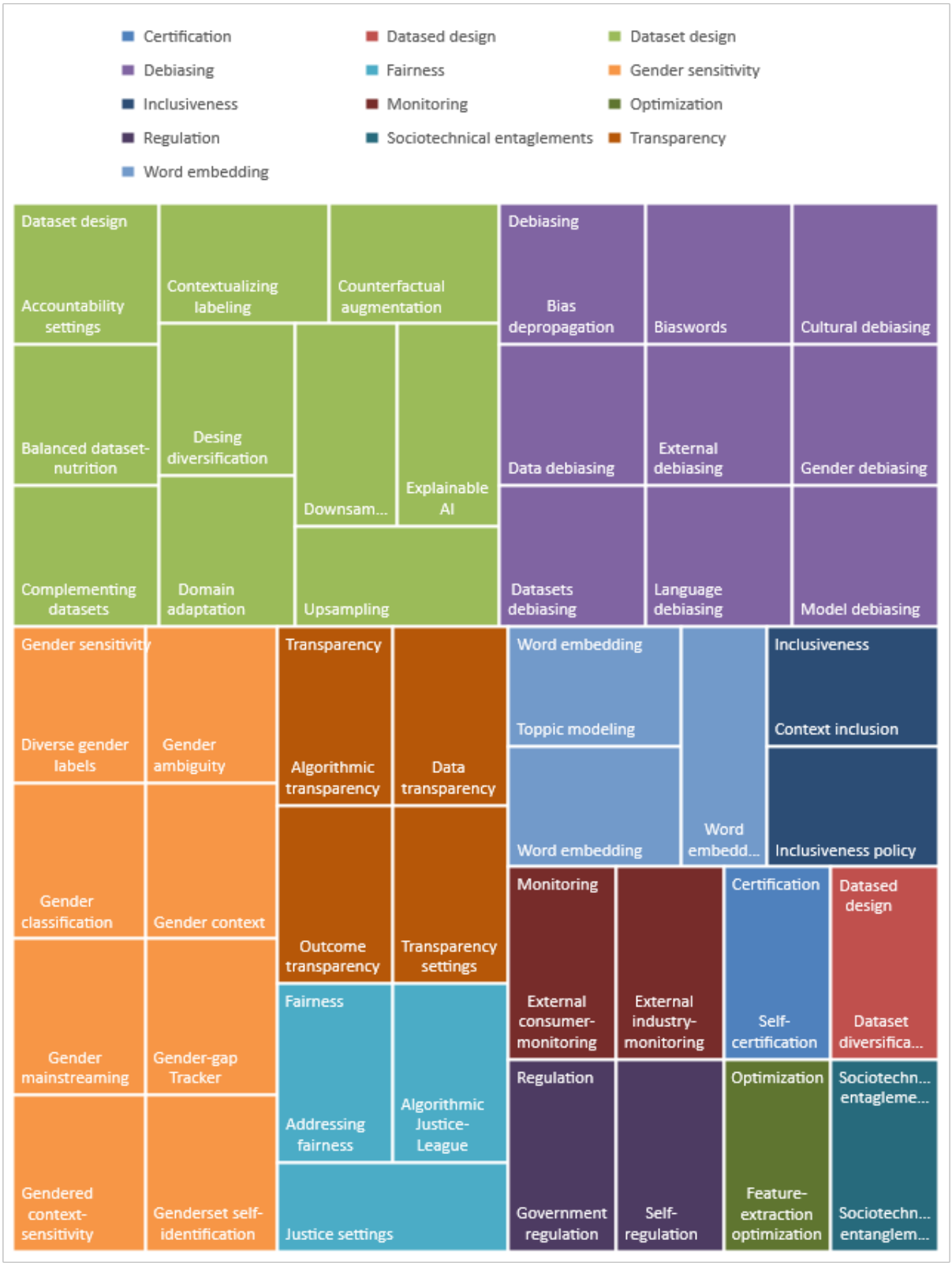

Figure 5. Treemaps with the leading strategies for overcoming (mitigating) gender biases in the studies related to artificial intelligence

http://doi.org/10.18080/jtde.v11n2.690

Word embedding and inclusiveness are represented with 14.29%, both categories found in two articles (n=2). Debating on biased language models and biased words, Kurpicz-Briki and Leoni (2020) reveal that, through language, our world is full of stereotypes; and they propose word embeddings as a solution to gender-biased AI. Similarly, Arseniev-Koehler *et al.* (2022) suggest word embedding to find solutions in avoiding gender biasing through topic modelling. Beside inclusiveness, Draude *et al.* (2020) also suggest sociotechnical entanglements, transparency, and fairness in elaborating different strategies of overcoming gender bias in AI. Finally, monitoring, certification and regulation solutions are found in Dwork and Minow's (2022) study, while optimization as a solution is recorded in Chen *et al.* (2022), all with 7.14% representation.

Conclusions

Theoretical implications

The results of this study have several theoretical implications. Regarding the categorizations created, fragmentation in the theme study was observed. Most authors on gender bias in AI have published only one article. This suggests the need for more in-depth research in the study area. Second, the survey of articles published by country revealed that most studies were conducted in the United States, highlighting the need for further studies in other developed and developing countries. Reflecting on this theme in other contexts may allow for greater depth of the subject. For example, of the 47 articles analyzed, only 14 presented causes of bias. Another 33 articles described only the existing types of bias without pointing out the reasons for them. Finally, this study contributes to the literature on gender bias and artificial intelligence by presenting and synthesizing concepts related to the types, causes, and ways of overcoming bias in AI by broad categories, contributing to research and researchers in the field.

Practical and managerial implications

This study has both practical and managerial implications. They are aimed at researchers in the field but also extend to developers of AI-related technologies, managers, women, and other stakeholders. For example, stakeholders can understand the types, causes, and ways of overcoming bias in AI, providing inputs capable of overcoming them. For example, a representative portion of the biases identified here could be addressed or diminished through initiatives such as audits conducted by managers. They can also create a kind of map or prioritization agenda for biases that need to be explored and mitigated, such as social and technical bias, individual bias, emergent bias, and linguistic bias. Managers should pay special attention to the extent to which social biases are most prominent in the literature.

http://doi.org/10.18080/jtde.v11n2.690

Simultaneously, technology developers can reflect on the topic, seeking to more accurately identify possible biases while developing AI-based tools and finding ways to overcome the problems that arise when using this research as a guide.

As for academics, such categorization allows mapping new research opportunities from its analysis and prioritizing biases, understanding them according to the highest incidence of studies. For example, a practical implication for academics would be to investigate the biases identified here and the possible associations or influences between them. Another practical implication for researchers is the exploration of factors that cause gender bias. Indeed, of the 47 articles that made up the final sample, only 14 highlighted the causes, denoting a theoretical and empirical gap that still needs to be explored by researchers in the area.

Among the leading causes emphasized, six groups stand out: systems, prejudice, culture, inequality, relationships, and interaction. Such groups also suggest that managers and public policymakers should better understand these aspects and act on the essence of their manifestation, such as prejudice. Different authors point out how, in essence, such prejudice is theoretically associated with social pressure for a considerably acceptable standard (Savoldi *et al.*, 2021; Huluba *et al.*, 2018; Vlasceanu & Amodio, 2022; Pair *et al.*, 2021; Petreski & Hashim, 2022; Schopmans & Cupac, 2021). However, how are these standards defined? Who stipulates them and their impact on systems? Such questions can help shed light on the cause's essence and help overcome it. Thus, it is hoped that this study can contribute to initiating reflection on this subject in the search for answers to these and other questions, as it identifies 58 overcoming strategies and the main types and causes of gender biases, representing a critical conceptual map from which managers and formulators can act.

Implications for capacity building

This study found that 88% of gender biases in AI were related to social biases, such as racial bias and male pattern bias, and the leading causes for these biases are social bias and the way data systems are designed. Thus, there is a need to encourage and support the creation of diverse teams with various perspectives, experiences, and backgrounds, including those from underrepresented groups, to work on all aspects and phases of AI projects. The findings of this study suggest, for example, the need to train software developers in the field of AI. Such training should primarily involve the main biases identified in this study.

Investing in and promoting the use of fairness and bias mitigation tools will help prevent and detect gender bias in AIs, building communities of practice and networks of experts in ethical AI and bias mitigation, including those explicitly focused on gender bias, encouraging public-private partnerships to support ethical AI capacity-building, including developing best

http://doi.org/10.18080/jtde.v11n2.690

practices, training programs, and tools for bias mitigation. Finally, supporting and funding research and innovation will advance our understanding of gender bias in AI and develop new strategies for addressing it.

Policy implications

Governments may need to regulate the development of AI models to ensure that they are free from bias and align themselves with existing legal and ethical standards. Policies must be implemented to ensure that the personal data used to train AI models are collected, processed, and used transparently and ethically, with appropriate safeguards against bias. They also need to be established to clarify the responsibilities and accountability of AI developers, organizations, and other stakeholders in preventing and overcoming gender bias in AI.

Limitations

This study has significant limitations. One connects to the inductive categorization of the data. Although such inductive categorization may be subject to other interpretations if performed by different researchers or be subject to the subjectivity of the authors of this study, it allows the identification and creation of new categories, which still need to be explored by the literature in the area. The second limitation is related to the scope of refinement, which is restricted to social and applied science studies. Therefore, sex biases related to medical or technological areas, among others, were not explored in the present study. Another limitation is the restriction of articles published in English, while excluding grey literature. Although such choices sought to filter the most relevant articles published in a more widely accessible language, they were simultaneously restricted by ignoring reflections, such as those published in congress articles and books.

Suggestions for future research

Future studies should address this study's limitations. In this sense, they could analyze the primary and secondary categorizations identified here, investigating, for example, whether they are mutually exclusive or whether there are still opportunities for new groupings among them. Simultaneously, new studies could advance into other unexplored opportunities. For example, studies have yet to be conducted in emerging and developing countries, especially in Africa and South America. Studying gender biases in artificial intelligence in more disadvantaged contexts could shed light on new biases that are still little explored or unidentified in developing contexts. Other opportunities for studies that are still poorly explored are associated, for example, with types of AI bias in gender studies. Although studies have mainly identified societal and technical biases, including different perspectives of individual and linguistic biases, they still need to be explored. For example, only some studies

http://doi.org/10.18080/jtde.v11n2.690

have addressed behavioural biases or their origins. Such studies are relevant because identifying the origin of biases can provide different practical suggestions to overcome them.

Another critical consideration in a future research study is the association between the research questions investigated in this study. This study sought to answer three fundamental research questions: 1) What are the main types of gender bias in AI? 2) What are the leading causes of these AI gender biases? 3) What are the main strategies for overcoming (mitigating) gender biases in AI? Further studies could, in turn, seek associations between more or less different RQs. For example, although this study has identified the main types of AI biases in gender studies, at the same time as the main strategies used to overcome them, it does not map or identify the strategies by type of AI bias. For example, are they recognizing overcoming strategies best suited to addressing societal bias? Are strategies that can be used for all biases denoting their relevance, or are they specific to each category? What is the association between AI bias and coping strategy? Is there any identifiable pattern between these two? These and other studies can broaden the understanding of the literature in this area and the corresponding policy and practical implications.

Concluding Remarks

Development and training with the deployment of gender-biased AI models may reinforce existing gender inequalities in society and contribute to their persistence over time, resulting in unfair or discriminatory decisions in various fields of social life (Deacon & Brooks, 1988; Kuppler, 2022). Gender bias in AI can also limit the ability of technologies to benefit all members of society, potentially leading to missed opportunities and unfulfilled potential for specific social groups (Deacon & Brooks, 1988; Breazeal & Brooks, 1997). It also raises ethical questions about the responsibility of AI developers and organizations to create and use these technologies (Fossa & Sucameli, 2022; Savoldi *et al.*, 2021). Therefore, it is essential to actively address and mitigate gender bias in AI to ensure that it benefits diverse social groups equally and fairly (Reyero Lobo *et al.*, 2022; Fyrvald, 2019; Nadeem *et al.*, 2022). This research can also serve as a theoretical foundation for systematizing and categorizing different types of bias, especially gender bias.

This study, contributing to reflections on gender bias in artificial intelligence, sought to answer three research questions: 1) What are the main types of gender bias in AI? 2) What are the leading causes of these AI gender biases? 3) What are the main strategies for overcoming (mitigating) gender biases in AI? Thus, we conclude that the main types of gender biases in AI are Societal, Technical, Individual, Emerging and Linguistic. Moreover, among the leading causes of gender-biased AI are Sociotechnical factors, followed by Societal and Technology systems. Finally, we highlight that the main strategies for overcoming (mitigating) gender-

http://doi.org/10.18080/jtde.v11n2.690

biased AI are related to: Dataset design, Debiasing, Gender sensitivity, Transparency, Fairness, Word embeddings, Inclusiveness, Monitoring, Regulation, Certification, Optimization and Sociotechnical entanglements.

References

Antony, J., Psomas, E., Garza-Reyes, J.A., & Hines, P. (2020). Practical implications and future research agenda of lean manufacturing: a systematic literature review. *Production Planning and Control, 32*(11), 889925. https://dx.doi.org/10.1080/09537287.2020.1776410

Arseniev-Koehler, A., Cochran, S. D., Mays, V. M., Chang, K. W., & Foster, J. G. (2022). Integrating topic modelling and word embedding to characterize violent deaths. *Proceedings of the National Academy of Sciences*, *119*(10), e2108801119. http://dx.doi.org/10.1073/pnas.2108801119

Asr, F. T., Mazraeh, M., Lopes, A., Gautam, V., Gonzales, J., Rao, P., & Taboada, M. (2021). The gender gap tracker: Using natural language processing to measure gender bias in media. *PloS one*, *16*(1), e0245533. http://dx.doi.org/10.1371/journal.pone.0245533

Bardhan, R., Sunikka-Blank, M., & Haque, A. N. (2019). Sentiment analysis as a tool for gender mainstreaming in slum rehabilitation housing management in Mumbai, India. *Habitat International*, *92*, 102040.http://dx.doi.org/10.1016/j.habitatint.2019.102040

Bhardwaj, R., Majumder, N., & Poria, S. (2021). Investigating gender bias in BERT. *Cognitive Computation*, *13*(4), 1008-1018. http://dx.doi.org/10.1007/s12559-021-09881-2

Breazeal, C., & Brooks, R. (1997). Gender Holes in Intelligent Technologies. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 1187–1192). IEEE.

Chen, X., Li, Z., Setlur, S., & Xu, W. (2022). Exploring racial and gender disparities in voice biometrics. *Scientific Reports*, *12*(1), 3723. https://doi.org/10.1038/s41598-022-06673-y

Conz, E., & Magnani, G. (2020). A dynamic perspective on the resilience of firms: A systematic literature review and a framework for future research. *European Management Journal*, *38*(3), 400–412. http://doi:10.1016/j.emj.2019.12.004

Corrêa, V. S., Brito, F. R. S., Lima, R. M., & Queiroz, M. M. (2022a). Female entrepreneurship in emerging and developing countries: A systematic literature review. *International Journal of Gender and Entrepreneurship, 14*(3), 300–322. http://doi:10.1108/IJGE-08-2021-0142

Corrêa, V. S., Lima, R. M., Brito, F. R. S., Machado, M. C., & Nassif, V. M. J. (2022b). Female entrepreneurship in emerging and developing countries: A systematic review of practical and policy implications and suggestions for new studies. *Journal of Entrepreneurship in Emerging Economies*. https://doi.org/10.1108/JEEE-04-2022-0115

Crawford, K. (2021). *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press (December 2022).

Crawford, K. (2013). The hidden biases of big data. *Harvard Business Review Blog*, Apr 1. Retrieved from http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data/. Accessed on Apr 10, 2023.

Das, S., & Paik, J. H. (2021). Context-sensitive gender inference of named entities in text. *Information Processing & Management, 58*(1), 102423. https://doi.org/10.1016/j.ipm.2020.102423

Deacon, T. W., & Brooks, D. R. (1988). Artificial Intelligence and the Bias of the Human Architect. In Proceedings of the 10th International Joint Conference on Artificial Intelligence (pp. 799–805). Morgan Kaufmann Publishers Inc.

DeFranza, D., Mishra, H., & Mishra, A. (2020). How language shapes prejudice against women: An examination across 45 world languages. *Journal of personality and social psychology*, *119*(1), 7. https://doi.org/10.1037/pspa0000188

Draude, C., Klumbyte, G., Lücking, P., & Treusch, P. (2020). Situated algorithms: a sociotechnical systemic approach to bias. *Online Information Review*, *44*(2), 325–342. http://dx.doi.org/10.1108/OIR-10-2018-0332

Dwork, C., & Minow, M. (2022). Distrust of Artificial Intelligence: Sources & Responses from Computer Science & Law. *Daedalus*, *151*(2), 309–321. https://doi.org/10.1162/daed_a_01918

Fyrvald, J. (2019). *Mitigating algorithmic bias in Artificial Intelligence systems*. Ph.D. Thesis, Uppsala Universitet. Available at https://www.diva-portal.org/smash/get/diva2:1334465/FULLTEXT01.pdf

Fossa, F., & Sucameli, I. (2022). Gender Bias and Conversational Agents: an ethical perspective on Social Robotics. *Science and Engineering Ethics*, *28*(3), 23. https://doi.org/10.1007/s11948-022-00376-3

Hägg, G., & Gabrielsson, J. (2020). A systematic literature review of the evolution of pedagogy in entrepreneurial education research. *International Journal of Entrepreneurial Behaviour and Research*, *26*(5), 829–861. https://doi.org/10.1108/IJEBR-04-2018-0272

Haraway, D. (1991). *Simians, Cyborgs, and Women: The Reinvention of Nature*. Routledge.

Haraway, D. (1987). A Manifesto for Cyborgs: Science, technology, and socialist feminism in the 1980s. *Australian Feminist Studies, 2*(4), 1–42.

Huluba, A. M., Kingdon, J., & McLaren, I. (2018). The UK Online Gender Audit 2018: A comprehensive audit of gender within the UK's online environment. *Heliyon*, *4*(12), e01001. https://doi.org/10.1016/j.heliyon.2018.e01001

Jones, J. J., Amin, M. R., Kim, J., & Skiena, S. (2020). Stereotypical gender associations in language have decreased over time. *Sociological Science*, *7*, 1–35. http://dx.doi.org/10.15195/v7.a1

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388–409. https://doi.org/10.1080/0960085X.2021.1927212

Kraus, S., Breier, M., & Dasí-Rodríguez, S. (2020). The art of crafting a systematic literature review in entrepreneurship research. *International Entrepreneurship and Management Journal*, *16*, 1023–1042. https://doi.org/10.1007/s11365-020-00635-4

Kuppler, M. (2022). Predicting the future impact of Computer Science researchers: Is there a gender bias? *Scientometrics*, *127*(11), 6695–6732. http://dx.doi.org/10.1007/s11192-022-04337-2

Kurpicz-Briki, M., & Leoni, T. (2021). A World Full of Stereotypes? Further Investigation on Origin and Gender Bias in Multi-Lingual Word Embeddings. *Frontiers in Big Data*, *4*, 625290. http://dx.doi.org/10.3389/fdata.2021.625290

Licklider, J. C., & Taylor, R. W. (1968). The computer as a communication device. *Science and Technology*, *76*(2), 13.

Machado, M. C., Vivaldini, M., & de Oliveira, O. J. (2020). Production and supply-chain as the basis for SMEs' environmental management development: A systematic literature review, *Journal of Cleaner Production*, *273*. https://doi.org/10.1016/j.jclepro.2020.123141

Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change, 175*. http://doi:10.1016/j.techfore.2021.121390

Martínez, C. D., García, P. D., & Sustaeta, P. N. (2020). Hidden Gender Bias in Big Data as Revealed Through Neural Networks: Man is to Woman as Work is to Mother? *Revista Española de Investigaciones Sociológicas (REIS)*, *172*(172), 41–76. https://doi.org/10.5477/cis/reis.172.41

Nadeem, A., Marjanovic, O., & Abedin, B. (2022). Gender bias in AI-based decision-making systems: a systematic literature review. *Australasian Journal of Information Systems*, *26*. https://doi.org/10.3127/ajis.v26i0.3835

Noble, S. U. (2018). Algorithms of oppression. In *Algorithms of oppression*. New York University Press. https://doi.org/10.18574/nyu/9781479833641.001.0001

Oldenziel, R. (1992). Cynthia Cockburn, Machinery of Dominance: Women, Men, and Technical Know-How (Book Review). *Technology and Culture*, *33*(1), 151.

Orgeira-Crespo, P., Míguez-Álvarez, C., Cuevas-Alonso, M., & Rivo-López, E. (2021). An analysis of unconscious gender bias in academic texts by means of a decision algorithm. *Plos one*, *16*(9), e0257903. https://doi.org/10.1371/journal.pone.0257903

Pair, E., Vicas, N., Weber, A. M., Meausoone, V., Zou, J., Njuguna, A., & Darmstadt, G. L. (2021). Quantification of Gender Bias and Sentiment Toward Political Leaders Over 20 Years of Kenyan News Using Natural Language Processing. *Frontiers in Psychology*, *12*, 712646. https://doi.org/10.3389/fpsyg.2021.712646

Paul, J., & Criado, A. R. (2020). The art of writing literature review: What do we know and what do we need to know? *International Business Review*, *29*(4), 101717. https://doi.org/10.1016/j.ibusrev.2020.101717

Patón-Romero, J. D., Vinuesa, R., Jaccheri, L., & Baldassarre, M. T. (2022). State of Gender Equality in and by Artificial Intelligence. *IADIS International Journal on Computer Science and Information Systems*, *17*(2), 31–48.

Petreski, D., & Hashim, I. C. (2022). Word embeddings are biased. But whose bias are they reflecting? *AI & Society*, 1–8. https://doi.org/10.1007/s00146-022-01443-w

Reyero Lobo, P., Daga, E., Alani, H., & Fernandez, M. (2022). Semantic Web technologies and bias in artificial intelligence: A systematic literature review. *Semantic Web* (Preprint), 1–26. https://doi.org/10.3233/SW-223041

Santos, S. C., & Neumeyer, X. (2021). Gender, poverty and entrepreneurship: A systematic literature review and future research agenda. *Journal of Developmental Entrepreneurship*, *26*(3). https://doi.org/10.1142/S1084946721500187

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, *9*, 845–874. https://doi.org/10.1162/tacl_a_00401

Scheuerman, M. K., Paul, J. M., & Brubaker, J. R. (2019). How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human*-Computer *Interaction*, *3*(CSCW), 133. https://doi.org/10.1145/3359246

Schopmans, H., & Cupać, J. (2021). Engines of patriarchy: ethical artificial intelligence in times of illiberal backlash politics. *Ethics & International Affairs*, *35*(3), 329–342. http://dx.doi.org/10.1017/S0892679421000356

Schwemmer, C., Knight, C., Bello-Pardo, E. D., Oklobdzija, S., Schoonvelde, M., & Lockhart, J. W. (2020). Diagnosing gender bias in image recognition systems. *Socius*, *6*, 2378023120967171. https://doi.org/10.1177/2378023120967171

Shrestha, S., & Das, S. (2022). Exploring gender biases in ML and AI academic research through systematic literature review. *Frontiers in Artificial Intelligence, 5*. https://doi.org/10.3389/frai.2022.976838

Tannenbaum, C., Ellis, R. P., Eyssel, F., Zou, J., & Schiebinger, L. (2019). Sex and gender analysis improves science and engineering. *Nature, 575*(7781), 137146. https://doi.org/10.1038/s41586-019-1657-6

Thelwall, M. (2018). Gender bias in machine learning for sentiment analysis. *Online Information Review*, *42*(3), 343–354. https://doi.org/10.1108/OIR-05-2017-0153

Tomalin, M., Byrne, B., Concannon, S., Saunders, D., & Ullmann, S. (2021). The practical ethics of bias reduction in machine translation: Why domain adaptation is better than data debiasing. *Ethics and Information Technology*, *23*, 419–433. http://dx.doi.org/10.1007/s10676-021-09583-1

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management*, *14*(3), 207–222. https://doi.org/10.1111/1467-8551.00375

Tubaro, P., & Coville, M., Le Ludec, C., & Casilli, A. A. (2022). Hidden inequalities: the gendered labour of women on micro-tasking platforms. *Internet Policy Review*, *11*(1). https://doi.org/10.14763/2022.1.1623

Turkle, S. (2005). *The second self: Computers and the human spirit*. MIT Press.

Vargas-Solar, G. (2022). Intersectional study of the gender gap in STEM through the identification of missing datasets about women: A multisided problem. *Applied Sciences*, *12*(12), 5813. https://doi.org/10.3390/app12125813

Vlasceanu, M., & Amodio, D. M. (2022). Propagation of societal gender inequality by internet search algorithms. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(29), e2204529119. https://doi.org/10.1073/pnas.2204529119

Waelen, R., & Wieczorek, M. (2022). The struggle for AI's recognition: understanding the normative implications of gender bias in AI with Honneth's theory of recognition. *Philosophy & Technology*, *35*(2), 53. https://doi.org/10.1007/s13347-022-00548-w

Wajcman, J. (2004). *TechnoFeminism*. Polity Press: Cambridge.

Wellner, G., & Rothman, T. (2020). Feminist AI: Can We Expect Our AI Systems to Become Feminist? *Philosophy & Technology*, *33*(2), 191–205. https://doi.org/10.1007/s13347-019-00352-z

Witherspoon, E. B., Schunn, C. D., Higashi, R. M., & Baehr, E. C. (2016). Gender, interest, and prior experience shape opportunities to learn programming in robotics competitions. *International Journal of STEM Education*, *3*, 1–12. https://doi.org/10.1186/s40594-016-0052-1