AI and Ethics
https://doi.org/10.1007/s43681-023-00386-2

REVIEW

A design perspective on how to tackle gender biases when developing AI-driven systems

Ana Santana González[1] · **Lucia Rampino**[1]

Received: 28 July 2023 / Accepted: 16 November 2023

Abstract
A growing awareness of bias in artificial intelligence (AI) systems has recently emerged, leading to an increased number of publications discussing ethics in AI. Nevertheless, the specific issue of gender bias remains under-discussed. How can design contribute to preventing the emergence of gender bias in AI-driven systems? To answer this question, we investigated the current state of AI ethical guidelines within the European Union. The results revealed that most guidelines do not acknowledge gender bias but address discrimination. This raised our concerns, as addressing multiple biases simultaneously might not effectively mitigate any of them due to their often-unconscious nature. Furthermore, our results revealed a lack of quantitative evidence supporting the effectiveness of bias prevention implementation methods and solutions. In conclusion, based on our analysis, we propose four recommendations for designing effective guidelines to tackle gender biases in AI. Moreover, we stress the central role of diversity in embedding the gender perspective from the beginning in any design activity.

Keywords AI-driven systems · Gender bias · Biased computer systems · Ethical guidelines · Gender diversity

1 Introduction

Over the last two decades, the many application fields of the design discipline (from products to communication, from interiors to services, from medical devices to fashion) have intersected with a revolutionary technology in constant development and expansion: artificial intelligence (AI). The design literature specifically addresses AI as *a* new design material [1, 2] or as a new design tool. In the first case, the final product is equipped with AI functionalities. In the second, AI is applied to enhance and optimize the outputs of the design process [3].

Beyond doubt, AI is significantly impacting design, bringing to many AI-driven experimentations. However, our guiding question in this article is *how can design impact AI-driven systems*? Indeed, when design enters in strict relationship with a new technology, it tends to modify it, introducing culture and the point of view of human beings [4].

1.1 The human responsibility in designing AI systems

In Science and Technology Studies, the idea of the sociotechnical system originated in the 1950s to overcome the technological determinism of the time. The aim was to emphasize reciprocity between humans and machines, in which mutual shaping of the social and technical systems always occurs [5]. AI systems can be understood as intelligent sociotechnical systems [6–8], comprising smart artifacts, human behavior, social arrangements, and meaning.

Similar to other sociotechnical systems, the design of AI systems involves choices regarding task allocation between humans and non-humans [9, 10]. To avoid the risk of falling into a new form of technological determinism, referred to as 'sociotechnical blindness' by Johnson and Verdicchio [11], it is vital to continually emphasize the central role of humans at every stage of AI system design and implementation: "*the behaviour of computational artefacts is in the control of the humans that design them*." [11, p. 584].

Acknowledging the role of people in creating, selecting, and providing data and instructions to AI systems can give rise to initiatives to prevent or limit negative impacts by acting on what is in total control of humans. Therefore, a systemic and holistic approach to developing and implementing

✉ Lucia Rampino
lucia.rampino@polimi.it

[1] Politecnico di Milano (IT), Milan, Italy

Published online: 15 January 2024

AI systems should be encouraged, with designers supporting engineers and developers in envisioning scenarios and guiding human and non-human behaviors [12].

1.2 Biased computer systems

In answering our initial question—how can design impact AI-driven systems?—we accepted the suggestion of Johnson and Verdicchio [11]. We, therefore, decided to address the problems in the human domains when a specific kind of undesired outcome emerges: gender bias.

Gender is a complex and multifaceted construct deeply embedded in society and culture. It encompasses not only biological characteristics, but also cultural norms, social roles and expectations. Gender bias refers to the unequal treatment or representation of individuals based on gender. It is a form of prejudice that can manifest in various ways (e.g., gender stereotypes), leading also to discrimination. Its consequences are pervasive in many domains and can significantly impact individuals and society. For instance, a study by Moss-Racusin et al. [13] found that science faculty members were more likely to hire and offer higher salaries to male candidates than equally qualified female candidates.

The presence of biased computer system is an issue debated since the end of last century:

> Computer systems, for instance, are comparatively inexpensive to disseminate, and thus, once developed, a biased system has the potential for widespread impact. If the system becomes a standard in the field, the bias becomes pervasive. If the system is complex, and most are, biases can remain hidden in the code, difficult to pinpoint or explicate, and not necessarily disclosed to users or their clients [14, p. 331].

The pervasive presence of AI can only make the issue of biased computer system more serious and urgent to be addressed.

The focus on gender bias was an authors' choice: being both women (an academic at a Polytechnic University and a young engineer in a research center), we are exposed to gender stereotypes that are deeply embedded in technological fields, still considered 'male territory' inside and outside academia. Our concern was confirmed by literature. According to Leavy [15], the over-representation of men in designing AI technologies could quietly undo decades of advances in gender equality, resulting in algorithms that perpetuate gender ideologies that disadvantage women.

Moreover, the studies conducted by Nass [16, 17] illustrated that the tendency to gender stereotype is so powerful that individuals also apply them to computers. Nass and Moon's article [17] is the origin of the so-called CASA (Computers Are Social Actors) paradigm, in which technological systems are deemed able to influence social relations and culture actively. Therefore, designers and developers' choices in designing AI systems can significantly affect how people interact and engage with them, ultimately shaping our social reality. Other more recent studies demonstrated the ability of AI systems to transmit the social concept of gender [18, 19], with the risk of perpetuating gender stereotypes.

Based on the above, we made our research question more specific: how can design contribute to prevent the emergence of gender bias in AI-driven systems? Our scope was to aid organizations in achieving practical solutions to gender bias, assuming a proactive attitude where the gender perspective is incorporated from the beginning.

1.3 The research methodology

To address our research question, we initially explored the topic of gender bias in AI, both in general and within the context of AI. Then a comprehensive literature review on AI ethics revealed a gap in in-depth analysis of gender bias in AI guidelines. To fill this gap, we focused our investigation exclusively on the European Union (EU), encompassing both EU institutions and member states. The choice to concentrate on this specific region is driven by the relatively limited attention it has received from other researchers, coupled with its proactivity in addressing AI ethics through regulations like the General Data Protection Regulation (GDPR) and the Artificial Intelligence Act. Furthermore, the EU comprises diverse member states with varying degrees of AI development and cultural perspectives. This diversity provides an opportunity to analyze how different countries address gender bias in AI.

We performed a review of any document published by EU institutions or member states in the AI Ethics Guidelines Global Inventory [20]. Ten documents were selected and analyzed through a framework developed to understand how current European guidelines address gender bias, also examining and comparing the proposed solutions. Based on this analysis, we defined four recommendations for designing effective guidelines to tackle gender biases in AI.

Even if—as also emerged from our analysis—significant overlap may exist in the approaches proposed to mitigate other kinds of unconscious biases that lead to discrimination (e.g., religious, and ethnic), we deemed it relevant to focus on a single kind of bias to elicit the multifaceted aspects behind it better. With slight modifications, our framework can be adopted to perform a similar analysis on any other form of bias.

The proposed framework and recommendations are our contribution to tackling the gender bias issue in AI systems. Such a contribution falls into the disciplinary field of design for policies, where design proposes itself as a transformation agent for organizational culture in the private and public sectors. Indeed, in recent years, design has evolved beyond the creation

of objects, systems and experiences to encompass critical discourse and decision-making in social, cultural, and ecological contexts [21]. This shift has been embraced by governments, fostering collaborations across public, private, community, and voluntary sectors for economic and social development [22]. Therefore, nowadays, one of the main design's jobs is to bring social awareness and action to complex issues ranging from poverty, education, and gender equity [23]. To perform this role, design equips itself with approaches, methods, tools, and frameworks to support pathways to systemic transformation towards a more equitable and sustainable society.

As Peters [24] highlights, effective policy design encompasses four key elements:

- Identifying the root cause of the problem (in our case, the origins of gender bias).
- Recognizing the available tools and resources (specifically in existing guidelines).
- Defining the intended goals (which, in our case, involve reducing hidden gender biases in AI design).
- Formulating an intervention plan. In our speculative discourse, this plan consists of four recommendations.

2 The origins of gender biases in AI

All human beings have biases and misconceptions [25]. In general, biases perpetuate inequalities, leading to the exclusion and disadvantage of individuals and groups that are already marginalized or vulnerable.

Regarding gender biases, Shields [26] found that one of the most remarkable similarities between girls and boys is that they are all knowledgeable about gendered stereotypes and that this understanding is already evident at young ages. However, this finding does not imply that biases are always conscious, i.e., deliberate, and that individuals are aware of them. On the contrary, biases are often implicit and unconscious, and can be influenced by factors such as past experiences, cultural stereotypes, and social conditioning. Unconscious biases are subtle and hard to detect. Therefore, it is fundamental for human beings to acknowledge them and their influence over their decisions.

In our literature review on gender bias in AI, we identified three main issues [15, 27–30]. Hereafter, each of them is outlined. A fourth issue will be discussed in the continuation of the article: the scarcity of specific standards and guidelines.

2.1 The permanence of gender stereotypes in the western world

Most feminists [31] believe that Western society is organized in a way that turns out to benefit men over women. This does neither imply that all men benefit equally from how society is structured since society oppresses men to different degrees, nor does it mean that all men participate in the system's continuation since men can oppose the oppression of other groups. Nevertheless, it indicates a general difference in how men and women are treated in society as a whole and in how they view themselves and others as gendered beings [32, 33].

As AI is built on human-generated data, it can inherit human biases, including those related to gender. In essence, AI technology reflects the biases entrenched in our culture, a complex challenge to address. What we can do in this regard is to rectify specific contributing factors to gender bias, addressing the lack of gender diversity and improving data quality.

2.2 Lack of gender diversity in AI development

The lack of gender diversity in the technology industry is one of the reasons which causes AI embedded products to be gender-biased, even if this often goes unnoticed [34–36]. This tendency is not only observed in western nations [37], but is a global practice [38].

Figure 1 illustrates the gender gap in technical positions in Europe in 2020. None of the 13 listed jobs have a higher female share. If there is a more significant male presence when designing an AI-driven system, it is more likely that such a system is conceived to suit and please this gender. Although this usually goes unnoticed, Intelligent Personal Assistants' default names (e.g., Alexa and Siri) and female voices are a good example of this. The designers selected this feature because female voices are perceived as 'supportive' and 'humble'. However, using feminine voices reinforces the stereotype of female servants and secretaries [34, 39].

2.3 The poor quality of training data

The sources used to train AI algorithms, including books, videos, audios, newspapers, and social media, often contain societal biases that the algorithms can learn and reproduce [27]. When biased sources are used, and no countermeasures are taken, the trained AI algorithm will eventually become biased too [41, 42]. However, this is a complex issue that we must not trivialize. A good example to outline such a complexity is the hiring domain, where audit studies spanning decades have revealed employers' discrimination against women and ethnic minorities [43–46]. Therefore, there has been rapidly growing interest in the use of algorithms in hiring, especially as a means to address or mitigate bias [47, 48], with proposed metrics to combat unfair practices [49]. However, other scholars [50] warn of data-driven algorithms potentially perpetuating inequalities under the guise of objectivity. A significant case is Amazon's recruitment algorithm which learned from

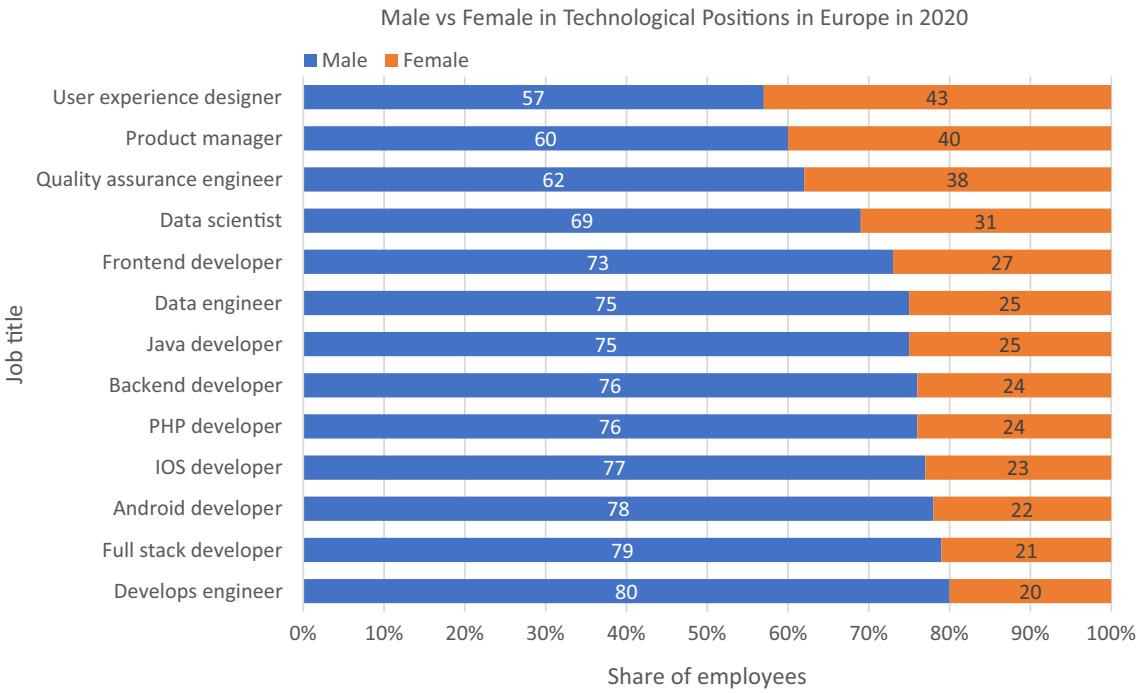

Fig. 1 Gender difference in technological positions in Europe in 2020 [40]

discriminative data and reproduced it by classifying male candidates as more suitable than female ones, even if this was not the case [51].

Other examples of possibly biased sources are provided by contextual word embedding models, such as ELMo and BERT that trained on large amounts of text data containing gender stereotypes, produce outputs that reflect such biases and can inadvertently reinforce existing stereotypes in downstream applications [52]. The same goes for AI systems trained on images. Testing commercial image recognition platforms, Schwemmer et al. [53] found out that images of women received three times more annotations related to physical appearance and that women in images are recognized at lower rates in comparison with men. According to the authors, these encoded biases affect the visibility of women and reinforce harmful gender stereotypes.

To prevent this all, ensuring good quality training data is one of the biggest challenges. The most frequently used tool is the implementation of guidelines and standards, which are the focus of our analysis. However, several studies anticipate that the number of guidelines addressing gender bias is minimal and that they do not always prove effective [34, 54, 55].

3 Reviewing the discourse on gender bias in AI

3.1 State-of-the-art literature review

To frame how scholars address the issue of gender bias in AI, we performed a state-of-the-art literature review, using the method by Webster & Watson [56] and reference-based-backward searches to identify the most influential papers [57]. The Webster & Watson method comprises the following steps [58]: (1) Data identification, (2) Data selection, (3) Classification and analysis and (4) Direction for future research.

We collected relevant literature using refined keywords; we searched multiple databases, including Google Scholar, the ACM digital library, and our university library. We included documents in English, Spanish, French, and Italian. The analysis of the total number of papers on AI released from 2010 to December 2022 in the ACM digital library and Google Scholar showed a significant increase in issued articles between 2018 and 2022 (see Fig. 2).

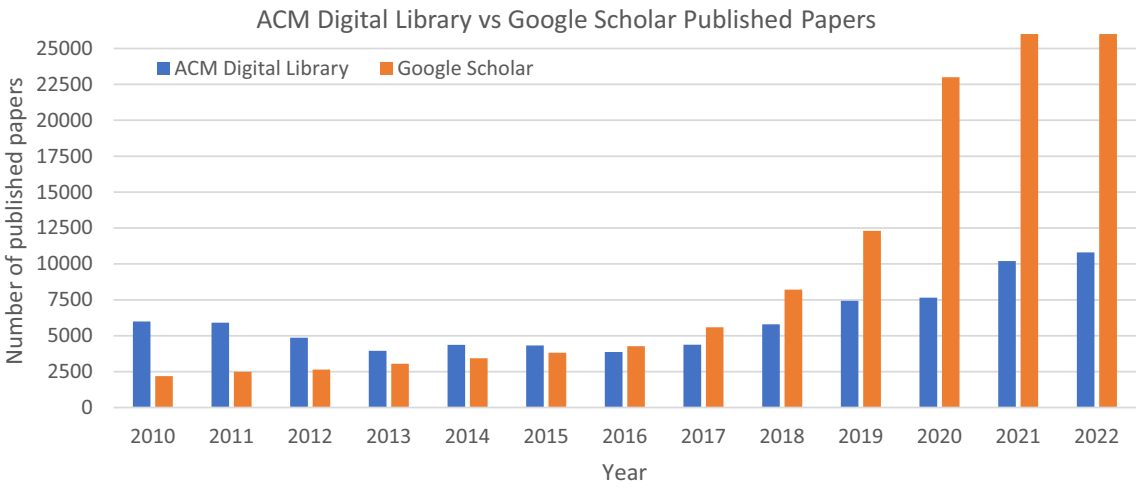

Fig. 2 Yearly quantity of published papers on AI

Therefore, our analysis primarily covered literature from this period. Nevertheless, we considered papers outside this timeframe when discovered during reference-based backward searches.

In step (1) of the Webster and Watson method, we began by selecting the keywords 'artificial intelligence' and 'gender stereotypes,' resulting in 11 papers published between 2018 and 2022. To expand our search, we introduced additional keywords such as 'gender bias,' 'ethical AI,' 'AI principles,' and 'AI guidelines' in combination with 'artificial intelligence.' This broadened our dataset to a total of 169 articles.

Moving on to step (2), we reviewed the titles and abstracts of these articles, selecting only the relevant ones for in-depth reading of the full text. Papers falling outside the scope of our research were excluded, resulting in 33 articles. We then conducted a reference-based backward search, which led to a final sample of 39 relevant documents.

In step (3), we focused on classifying and analyzing these 39 articles, as described in the 'The European Landscape of Ethical AI' section. This analysis was instrumental in developing our framework, which we used to analyze the 10 selected guidelines.

Finally, in step (4), we synthesized the outcomes of our state-of-the-art review with insights from the guideline analysis, as described in the following paragraph. This synthesis resulted in one key insight and four recommendations for designing guidelines to address gender biases in AI.

Many of the 39 documents relevant to our research revolved around guidelines and standards for ethical AI. Indeed, the efforts of companies and organizations to become more ethically friendly have brought an outbreak in the number of AI standards and guidelines issued by nations, research institutions, and private companies in the past few years. This has triggered the interest of many scholars who compared policies issued in the European Union, China, the United States, and the United Kingdom.

Thinyane and Goldkind [59] report that, until 2020, over 160 AI-related guidelines were published worldwide, encouraging researchers to examine their contents and effectiveness to uncover how organizations envision the future of AI. Scholars identified specific trends, such as the similitude in the number of issued documents by both public and private entities [54, 60, 61], the lack of discussion on the topic of gender diversity within the AI community [55] and the need to accompany guideline documents with additional tools [55, 62].

3.2 Guidelines for ethical AI review

Considering the above, we deemed it relevant to analyze the AI guidelines landscape, to understand if and how existing guidelines address the gender-bias issue. Indeed, none of the analyzed 39 papers delved deeply into the topic of gender bias in AI guidelines. Moreover, in terms of geographical distribution, most of the analyzed paper either presented a global perspective on AI guidelines or focused solely on the United States. As said, the EU demonstrated proactivity in addressing the topic of AI ethics, therefore we made the decision to exclusively evaluate documents issued within the European Union. Our search in the *AI Ethics Guidelines Global Inventory* [20] yielded a total of 52 articles from European Union institutions or member states, surpassing the numbers for the United States (44), China (4), and the United Kingdom (19). Within the 27 EU member states, only 12 have published such documents, as depicted in Fig. 3.

After applying a language filter, we narrowed our focus to 35 guidelines published between 2014 and 2021. Figure 4

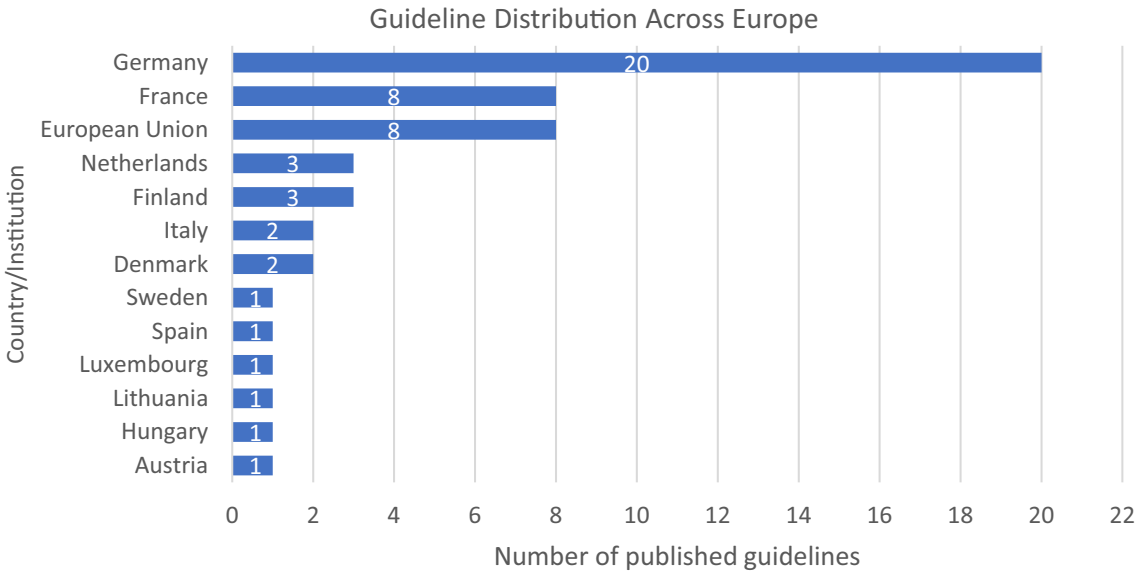

Fig. 3 Published guideline distribution from different EU member states

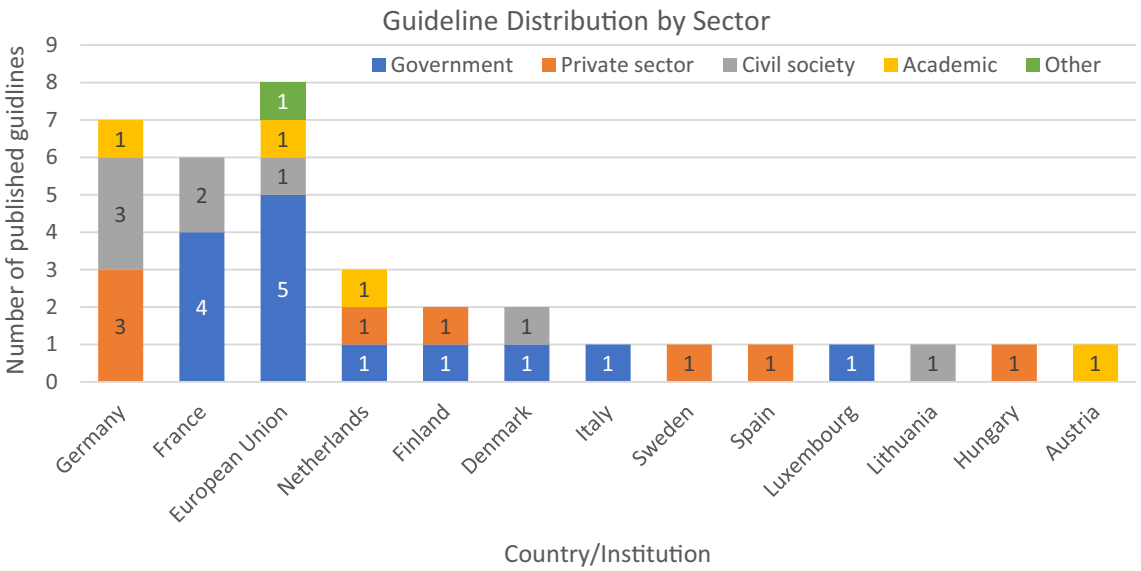

Fig. 4 Guideline distribution by sector and country

illustrates that most of these guidelines were issued by Governments, civil societies, and private businesses. These findings align with prior research [54, 60, 61]. Notably, Hagendorff [55] emphasizes the importance of government involvement in ethical guideline development, as private companies often establish principles that primarily serve their economic interests rather than the broader societal good.

After individually reviewing these 35 guidelines, we conducted an in-depth analysis of ten, as presented in Table 1. We excluded 18 documents that did not align with the purpose of our analysis (e.g., declarations of trust or European Union parliamentary meeting minutes). Additionally, seven were omitted for addressing discrimination and gender bias in a cursory manner, often limited to two-sentence discussions. Among these, a few offered potential solutions to AI-related discrimination. Nevertheless, the ten selected guidelines were chosen for their superior quality, meeting criteria encompassing depth, relevance, and comprehensiveness in addressing bias prevention, not exclusively limited to gender bias.

Table 1 Guidelines selected for the analysis

Document name	Issuer	Date	Sector	Issuing country
SAP's guiding principles for artificial intelligence	SAP's AI Ethics Steering Committee	2021	Private	Germany
Digital Ethics—A guide for professionals of the digital age	Club Informatique des Grandes Entreprises Françaises (CIGREF)	2018	Civil Society	France
For a meaningful artificial intelligence towards a French and European Strategy (Mission Villani)	French Strategy for Artificial Intelligence	2018	Government	France
Artificial intelligence and data protection	Council of Europe	2018	Government	Europe
Ethics Guidelines for Trustworthy AI	High Level Expert Group on Artificial Intelligence	2019	Government	Europe
A framework for the ethical use of advanced data science methods in the humanitarian sector	The humanitarian data science and ethics group	2020	Academic	Europe
Data ethics decision aid (DEDA)	Utrecht university	2017	Academic	Netherlands
Data for the benefit of the people	Danish expert group on data ethics	2018	Government	Denmark
Principios de IA de Telefónica	Telefonica	2018	Private	Spain
AI UX: 7 Principles of Designing Good AI Products	UX studio team	2018	Private	Hungary

The results of our literature review and guidelines analysis are described in the following sections.

4 The European landscape of ethical AI

Guidelines can be seen as rules, principles, or recommendations, depending on the content or the issuer. Some argue that they are ineffective, while others (cited in [54]) state that they have proven to influence decision-making in specific fields.

Benjamins [63] suggests that companies and organizations create guidelines for two reasons: to decide what principles to adopt, and to introduce a methodology for implementing the principles adopted. Others [54] believe that certain issuers worry more about their moral obligation than addressing potential ethical pitfalls.

The literature we analyzed in our state-of-the-art review tends to compare specific guidelines to determine which documents cover more AI-related issues. Non-discrimination and bias prevention emerge as the most addressed matters. Some researchers have focused on describing methods for companies to identify and successfully implement these issues. For instance, Benjamins [63] proposed a three-step approach for businesses committed to the responsible use of AI but lacking knowledge and experience on it.

Whittlestone et al. [62] discussed the limitations of AI principles, defining them as too broad and ineffective to benefit society. They suggest that principles should be transformed and formalized into guidelines, standards, and regulations. In addition, they agree with other researchers [55] on the necessity of accompanying this documentation with additional tools useful in situations where conflicts may arise. The researchers give several examples where principles come into conflict in practice and then state that entities should focus on solving these tensions, which will result in the development of useful and relevant frameworks.

Dignum [64] investigated the current state of ethical AI by analyzing guidelines that address ethics and design from three different perspectives: 'by design', 'in design' and 'for design'.

- Ethics by design: the technical integration of ethical values and principles in AI systems.
- Ethics in design: how entities ensure that AI development processes are aligned with ethical principles.
- Ethics for design: the effectiveness of the implemented codes of conduct and standards for ensuring developers' integrity as they design, develop, employ, and manage AI systems.

The findings of her study have proven to be successful for companies and organizations to design and implement methods that consider all ethical aspects of the design process.

Wagner [65] states that for ethical approaches to be taken seriously, they should count on external participation and provide a mechanism for independent inspections. To allow for this, transparent explanations of why decisions were taken should be available.

Benjamins et al. [66] described how Telefónica, a multinational company that develops and uses AI, has implemented specific tools to minimize the risk of undesired consequences. Sharing the company's experience is intended to encourage and help other organizations promote and implement good AI practices.

The recent increase in the number and variety of guidance documents confirms the growing interest in AI ethics. Additionally, the similar number of such documents issued by both public and private entities underscores the significance of AI ethics for both sectors [54, 60, 61]. However, the content within these documents varies significantly.

Nevertheless, some recurrent themes can be identified, such as privacy, accountability, fairness, transparency, and explainability.

In terms of geographical distribution, the previously mentioned focus on the United States in particular, and on Western countries in general, highlights an unequal participation in the AI ethics debate, with an evident underrepresentation of geographic areas such as Africa, South and Central America, and Central Asia. Jobin et al. [54] express concerns about this uneven distribution of issued documentation. They argue that this produces an unevenness of power, leading to more economically developed countries shaping this debate, while local knowledge, cultural pluralism, and global fairness demands are neglected.

Finally, there is a lack of quantitative research on proving the effectiveness of implementing principles, guidelines, and toolkits in the private and public sector.

4.1 The analysis' framework

To what extent have EU companies and organizations implemented standards and tools to prevent gender bias in AI systems? We designed a framework for our guidelines analysis (see the detailed breakdown in the Appendix), based on findings from our literature review (refer to Table 2 for details).

The framework was aimed to determine the following:

- The level of acknowledgment of gender biases.
- The presence of preferred solutions by entities operating in specific sectors.
- The level of detail at which solutions have been implemented in entities.

Unlike other approaches, our framework uses 'Yes' or 'No' questions for analysis, making it easier to quantify and compare the results. It consists of three sections addressing (i) the problem, (ii) the actions, and (iii) the results.

To effectively address a challenge, the initial imperative is to recognize and confront its presence [67]. Therefore, the first section of our framework was designed to determine whether the guideline issuer recognizes gender bias as a problem and is aware and capable of understanding its consequences. Questions embedded in this section are intended to first give a broad perspective on the use of AI within the entity: '*Does the document state where/how the AI technology is used within the company or organization?', 'Do they develop the AI or integrate it their product/service?'* or *'Who is the product/service targeted at?'*. The following questions specifically address the issue of gender bias, *'Is gender bias identified and explained as an issue under the fairness and discrimination principle',* and also *'Does the issuer explain the potential cause/origin of gender discrimination within their company or organization?'.*

The second section of the framework focuses on the actions, i.e., the solutions and implementation strategies proposed to counter-fight gender biases in AI. To the best of our knowledge, there is no existing analysis regarding the extent to which this AI-related issue has been addressed. However, we have drawn from insights in some reviewed articles to develop this section. Notably, researchers such as those in [55, 62, 64] agree that checkbox guidelines are insufficient. Therefore, we have included questions such as '*Are any specific solutions proposed?'*. Additionally, to assess the level of ethics within the design process, we have included the question '*Will there be any consequences if the proposed solutions are not followed/used?'* to cover the ethics *by design* perspective. The question '*Is there a technical explanation (code) provided?'* addresses the ethics *in design* perspective [64].

Regarding solution implementation methods, the analysis aimed not to find the optimum procedure like other

Table 2 Overview on the state-of-the-art on AI implementation methodologies recommendations

Author	Key recommendations
Benjamins (2020)	There should be a distinction between principles relevant to governments and those targeted to companies
Benjamins, Barbado and Sierra (2019)	The methodology was designed by multi-functional departments
	Several solutions were developed to reduce potential risks of AI
	Training programs were extended across the company and adapted to different departments
Dignum (2018)	Guidelines should address ethics and design from different perspectives (ethics by/in/for Design)
Wagner (2018)	Ethical approaches should count with external participation
	Transparent explanations on why decisions were taken is crucial
	Companies and organizations should develop guidelines which focus on issues which recur in their businesses activity or in their sector
Whittlestone et al. (2019)	Principles are too broad and not enough to produce an effect that will benefit society
	Guidelines should be accompanied with specific tools
	Focusing on specific conflicts will help to reduce the current gap between principles and practice

researchers [54], but to understand the level of engagement of companies and organizations in the action stage. We addressed this by incorporating the following questions '*Does the issuer present an implementation plan/approach for the proposed solutions*?', '*Will everyone in the organization or business be affected by this new methodology in the same way*?' and '*Are the proposed solutions meant to be used by individuals or collectives*?'.

The last section of the framework is intended to investigate the success of the solutions proposed, covering ethics *for design* [64]. To this end, three questions were included: '*Has feedback been received from the different stakeholders involved in the process?*' '*Has any literature been published on the proposed methodology?*' and '*Are other businesses or organizations using the same approach?*'.

4.2 Results of the guidelines' analysis

For each of the ten selected guidelines, we completed a framework form to assess the level of detail at which the issue of gender bias is addressed and the success of proposed solutions. Table 3 summarizes the findings, showing, for each guideline, the presence of the following aspects:

- A description of where and how AI is used within the company or organization.
- Identification of gender bias as a specific issue.
- Solution(s) to prevent gender bias or discrimination.
- An implementation plan.
- Proof that the proposed solutions have been successful/unsuccessful.

In Table 3, when a guideline discusses any of the mentioned points, the corresponding cell is marked with an 'X.' Notably, our analysis reveals that not even half of the guidelines address more than half of these aspects, and none address all five.

From the queries under Sect. 1 of the framework—*Problem Identification*—we found that only two documents explicitly stated where or how AI technology is used within their organization. Even more striking is that only one of the analyzed guidelines—'*For a Meaningful Artificial Intelligence Towards a French and European Strategy*' (Mission Villani) by the French Strategy for Artificial Intelligence (FSAI)—considers gender bias as a specific issue. The FSAI recognizes that the potential origin of gender bias is poor data quality and pre-existing bias in society. Therefore, in their report, they discuss inclusiveness and diversity in the technological sector in France, proposing several solutions to the reduced number of females studying computer science, the limited representation of female engineers in the digital industry and within executive committee, and the gender pay gap. For example, they suggest initiatives like teaching coding to girls and setting targets for female enrollment in digital courses.

The remaining nine guidelines take a broader approach, addressing multiple biases at once, including gender and religious bias, under-representation of minorities, and ethnic discrimination. However, addressing multiple biases simultaneously raises some concerns, as it may not effectively tackle any of them due to their often-unconscious nature. As a result, specific issues like gender bias potentially remain

Table 3 Summary of findings

Issuer	Where/how AI is used within the company or organization	Gender bias identified as a specific issue	Solution (s) to prevent gender bias or discrimination	Implementation plan	Proof that the proposed solutions have been successful/unsuccessful
SAP's AI ethics steering committee			X		
CIGREF			X		X
French strategy for artificial intelligence		X	X		
Council of Europe	X		X		X
High level expert group on artificial intelligence			X		
The humanitarian data science and ethics group			X		
Utrecht university			X	X	X
Danish expert group on data ethics			X	X	
Telefonica	X		X	X	X
UX studio team			X		

Table 4 Most common solutions

Issuer	Checklist/questionnaire	Training	Bias prevention algorithm	Diversity team or ethics committee	Other
SAP's AI ethics steering committee				X	
CIGREF	X	X			X
French strategy for artificial intelligence		X		X	X
Council of Europe	X			X	
High level expert group on artificial intelligence	X	X	X	X	X
The humanitarian data science and ethics group	X		X	X	X
Utrecht university	X				
Danish expert group on data ethics					X
Telefonica	X	X	X		
UX studio team					X

underdeveloped. It is indeed challenging to propose accurate solutions if biases are not clearly identified.

Nevertheless, all ten guidelines propose at least one solution to mitigate bias and discrimination. Table 4 provides insights into the tools and resources most recommended.

Overall, our analysis underscores the need for more targeted efforts to address specific biases and calls for action from designers to develop targeted tools for addressing these issues.

The second section of the framework, labeled 'Action Implementation', revealed that just three out of ten documents presented an approach to implement the proposed solutions (refer to Table 3 for details). The depth and specificity of these implementation plans vary, largely due to differences in sectors.

On the one hand, Utrecht University developed the *Data Ethics Decision Aid* (DEDA), a toolkit designed to facilitate initial brainstorming sessions aimed at mapping ethical issues in data projects [68]. This toolkit shows that this institution believes that businesses should establish ethical guidelines not only for specific projects but also for their daily activities.

On the other hand, the guidelines of Telefónica—a private company—are very specific on how to implement ethical solutions and who would be affected by them. Their approach begins with defining values and boundaries. Subsequently, they offer both technical and non-technical online training to all employees, with the length and duration of training varying based on employee profiles. An online checklist is also in place for workers involved in creating AI. Additionally, they offer various technical tools, developed both internally and externally, to mitigate potential issues. Lastly, a governance model is integrated to assign responsibilities and escalate concerns that developers cannot resolve using the aforementioned tools.

Regarding the target audience, these two examples represent a contrast between a generic and a 'custom-made' approach, illustrating how complex implementing a solution can be. DEDA is designed for businesses and organizations, even if their implementation approach might seem too general. However, it can be adjusted and personalized for any company's needs. Telefónica's implementation approach, conversely, was specifically tailored to the company's activities. While other businesses can follow a similar plan, some adjustments may be necessary.

Out of the ten documents, only four state that the proposed solutions were 'successful.' In assessing the evidence behind such a declaration, we considered the specific nature of the issuer.

The documents released by CIGREF [69] and the Council of Europe [70] highlight ongoing campaigns that align with the solutions they have proposed. Given that the issuers operate in the civil society and government sectors, respectively, this approach is considered acceptable. Indeed, guidelines issued by governing bodies are typically broader, intended to guide other organizations to adapt them to their day-to-day activities.

On the other hand, Telefónica's approach was considered successful for two reasons. First, they stated that the proposed approach was adapted from other methods that have already proven to be effective. Second, the company has published various papers [63, 66] explaining and discussing this matter.

In the case of other private companies, it is surprising that there is no mention of whether different stakeholders within the business have provided feedback on the implemented solutions. This is a potential pitfall since ensuring stakeholders' participation in all stages is crucial to keep them motivated and encourage them to use bias mitigation tools.

Worth noting is that while four issuers presented proof of implementation approaches or solutions being successful, none of these have been quantitatively measured. Such measurements could provide a more robust evaluation of success in addressing bias and discrimination.

Table 4 provides insights into the tools and resources most recommended in the ten analyzed guidelines. Checklists and questionnaires are the most common solutions. Notably, five of the six guidelines that proposed implementing checklists or questionnaires also suggested at least another solution. This aligns with published findings on the need to accompany checklists with additional resources to prevent (gender) bias.

The next most popular solution recommended in the guidelines is the creation of diversity teams or ethics committees, followed by training courses and the implementation of bias prevention algorithms. It is worth noting that none of the documents proposing this last solution provided a technical explanation of the algorithm to be implemented.

The results did not reveal any specific trend as to whether certain sectors preferred some solutions or others. Checklists, diversity committees, and training courses are the most common solutions. However, many issuers proposed additional solutions that were reviewed and grouped.

Several documents called for creating auditing platforms or committees, others for introducing ethical certifications such as those submitted for products made from recycled materials. Guidelines such as *DEDA, HDSEG*, and *DEGDE*, which focus on data ethics, suggest creating data custodian departments to control the type of training data provided to the engineers. Finally, other propositions involved raising incentives to conduct more research regarding ethical AI and allowing for users' feedback. This means testing algorithms with real users and incorporating features to enable users to provide inputs and feedback to be used as training data.

5 Four recommendations to tackle gender biases in AI

AI is a new and complex matter; thus, it is not surprising that there is still a small number of guidelines available and that—as our analysis shows—only few companies or organizations consider the overall spectrum of ethics *by/in* and *for* design. The AI industry is in its early stages, and it's currently exploring the most effective approaches and solutions. Additionally, we acknowledge the perspective put forth by scholars who suggest that some documents might be issued as part of a marketing strategy to showcase an organization's commitment.

Determining the key elements of an effective guideline can be challenging, given the significant variation in document content, depending on the sector, target audience, and EU country. Nevertheless, our state-of-the-art literature review and analysis of the ten guidelines reveal one crucial insight and four recommendations. All of them point to the need for governments and other entities to work collaboratively, providing each other with continuous feedback.

The insight is that guidelines should be structured along two dimensions: general principles for AI ethics and specific issues falling under those principles. In practical terms, governments should define overarching principles like fairness and non-discrimination, while companies should outline their approaches to specific issues, like gender bias, within these broader principles.

To incentivize companies and organizations to innovate and find new solutions, governments can consider offering economic incentives or EU-approved ethical certifications, such as 'non-gender-biased approval seals', to companies that successfully develop or implement solutions. Another interesting initiative proposed by the DEGDE [71] is making companies declare their data ethics policies as part of their annual financial statement. To adapt such an initiative to gender bias mitigation, companies can be asked to show their employees' diversity level.

Moreover, the following four recommendations are designed to enhance guidelines addressing gender biases.

1. Proposed solutions should be specific to a single issue, not a principle. Moreover, solutions must be more than a simple checklist.

To better understand the relationship between principles and issues in AI guidelines, we can draw an analogy with books. In this analogy, principles can be likened to chapters in books, while specific matters correspond to subsections. Just as chapters in a book are distinct but related sections, the issues or subsections under a principle are not necessarily interrelated, but they all fall under the umbrella of that principle.

One of the most common principles we have observed in literature is fairness and non-discrimination. As highlighted in the report by Fjeld et al. [61], specific issues must be addressed under this principle, including bias prevention, the use of representative and high-quality data, promoting equality, and ensuring inclusiveness in design. Proposing a non-discrimination questionnaire that engineers must fill in while developing an algorithm would possibly prevent bias. However, addressing issues related to poor-quality data or achieving equality requires additional resources, such as measures to ensure equal pay or the establishment of diversity committees.

It is worth noting that most of the solutions presented in the analyzed guidelines have been developed to address discrimination in general, and there are relatively few design tools specifically created for gender bias prevention. This underscores the need for action on the part of designers to develop tools tailored for addressing gender bias. It is also worth considering that some tools designed to prevent discrimination could be adapted for gender bias. However, it is important to recognize that many researchers have demonstrated that using a single tool (above all if this tool is a checklist) may be insufficient to comprehensively solve

the problem (for further insights on this point, see the third recommendation).

2. Solutions and implementation methods should be both transparent and adapted to the level of involvement of employees with AI technology.

Transparency is critical when implementing a new method, especially in the context of AI and ethics. In companies, ensuring transparency means that everyone within the organization should be well-informed about the issues related to AI ethics and understand how the proposed solutions are intended to address these concerns. Furthermore, achieving transparency also entails recognizing the varying levels of involvement and knowledge different departments or individuals may have with AI technology. Some departments or team members may already possess a strong understanding of AI and its ethical implications, while others may be relatively new to the field. Tailoring transparency efforts to match the specific needs and knowledge levels of each department or individual can help ensure that everyone is on the same page, fostering a more cohesive and informed approach to AI ethics.

3. The proposed solutions should be tested with stakeholders and allow for feedback.

Solutions should be validated in advance through engagement with the affected stakeholders, encouraging an ongoing feedback loop to foster a culture of continuous improvement. The success of these solutions should be quantified, and the results shared with relevant stakeholders to ensure accountability and build of trust. This includes measuring the reduction in gender bias compared to the absence of these solutions or testing and comparing different approaches, as demonstrated by CIGREF [69] in their guidelines.

In line with [65], it is recommended that all private companies conduct both external and, where feasible, internal audits focused on AI ethics. These audits play a crucial role in maintaining control over ethical practices and verifying that prevention procedures are consistently adhered to.

Additionally, as emphasized by several issuers, including SAP [72], the Council of Europe, the HLEG [73], and Telefónica [74], ethics or diversity committees should be created to address complex issues that may not be fully resolved through the provision of tools to designers and engineers.

Telefónica's approach serves as a model for effective implementation. However, it is essential to stress the importance of enabling and facilitating feedback mechanisms in these processes.

4. Guidelines issued by governments should include a list of successful initiatives.

Government-issued guidelines should not only outline the principles and best practices for addressing gender bias and discrimination in AI, but also include a list of existing initiatives that have already proven successful. Such a compilation can serve as a reference point for all companies seeking to implement ethical AI practices effectively. By highlighting initiatives that have demonstrated their value, these guidelines can guide companies toward adopting strategies that have a track record of success. Notably, some forward-thinking guidelines, such as those issued by CIGREF and the Council of Europe, already include this valuable resource. Their inclusion of successful initiatives sets an exemplary standard for sharing best practices within the AI ethics domain, ultimately benefiting a wide array of organizations in their quest to combat gender bias and discrimination.

5.1 The central role of gender diversity

Unlike other issuers that state that gender bias or discrimination is solved by providing engineers with good-quality training data, the FSAI discusses other topics such as gender equality and diversity. Our analysis highlighted the recurrence of the issue of gender diversity as a crucial indicator that could drastically reduce gender bias in AI systems.

As [55] pointed out, the discourse of AI ethics is primarily shaped by men, which explains the limited consideration of gender bias as a specific issue in guidelines, as confirmed by our analysis. Therefore, it is crucial to promote gender diversity in technological companies and organizations to break the cycle that perpetuates gender bias.

Figure 5 summarizes the four key issues driving gender stereotypes in AI, as discussed in this article.

Diversity is positioned at the top of the diagram, serving as the catalyst for solutions that influence the subsequent issues. Data are placed at the bottom of the diagram, representing the final step toward achieving unbiased systems. As mentioned earlier, while data plays a significant role in AI ethics, addressing other factors is essential before attaining high-quality data.

Fostering diversity within a company, university, or organization can have a profound impact on its culture and

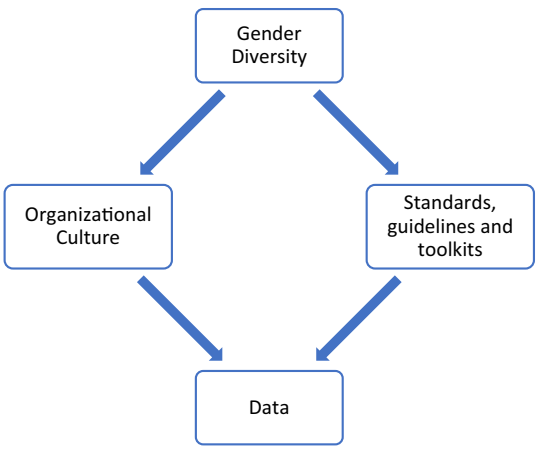

Fig. 5 How diversity can positively impact other factors affecting gender bias in AI

decrease biases, directly influencing the quality of data used to train AI algorithms. Moreover, promoting diversity within such entities is instrumental in shaping robust ethical guidelines and standards. This approach ensures that a broader spectrum of perspectives, including those of various groups and minorities, is considered, ultimately leading to the development of more effective mechanisms for detecting and preventing subpar training data in algorithms. The culmination of these improvements results in a significantly higher success rate in preventing bias during the AI design process.

Many AI experts [15, 29, 41, 55] have analyzed different ethical issues within AI, yet they all agree that gender *diversity* is the key factor to avoiding the development of discriminative algorithms.

Some may argue that fostering diversity into their organization requires time and resources. However, numerous studies have shown that diversification offers multiple benefits. For example, a Boston Consulting Group study [75] found that companies, especially in the technology sector, with more diverse teams experience a 19% increase in revenues due to innovation. Additionally, businesses with more women tend to be more profitable. A Pew Research Center survey [75] indicates that females are 34% more likely to exhibit honesty and ethics, 30% more likely to provide fair pay and benefits, and 34% more adept at finding compromises.

In the long term, governments should collaborate with educational institutions to introduce ethics education from an early age [76]. Schools can host workshops to inspire young girls to pursue engineering and technological fields, offering insights into the exciting aspects of tech careers and providing role models. Replicating successful coding education campaigns from several Asian countries can be beneficial. European technical universities should also intensify efforts to achieve gender balance among both the students' cohort and in the academic staff.

It is essential for companies, organizations, universities, and governments to incorporate diversity within their teams, setting in motion a chain of positive effects to combat gender bias in AI systems. However, it is crucial to acknowledge that this is a long-term process, requiring several years to witness substantial results. Cultural change takes time, and resistance to change is inherent to human nature [77].

6 Conclusion

We explored how design can combat gender bias in AI systems and formulated four recommendations based on a literature review and an assessment of ten EU AI ethical guidelines. Most guidelines lack details on how AI is used within the organization and do not explicitly address gender bias specifically but discrimination more broadly. The most common solutions to discrimination include checklists, questionnaires, diversity or ethics committees, and employee training. Quantitative proof of implementation success, solution effectiveness, and user feedback consideration is often lacking.

Gender diversity is identified as a crucial issue from our analysis. Promoting diversity within organizations improves culture, reduces societal biases, positively impacts data used for AI training, and results in more accurate guidelines and standards. This, in turn, leads to better mechanisms for detecting and preventing poor-quality training data, ultimately enhancing gender bias prevention in AI design.

Our analysis framework and recommendations contribute to addressing gender bias in AI, falling within the realm of design for policies, where design plays a transformative role in organizational culture.

6.1 Limitations of the study and future work

This article has shed light on the current state of gender bias within European guidelines, supported by relevant literature wherever possible. However, it is important to acknowledge the limitations of this investigation. Notably, the analysis was confined to guidelines from a single database [20], which may not capture all relevant policies published in other databases or on issuer websites.

Furthermore, while we have proposed recommendations for comprehensive guidelines and highlighted the significance of gender diversity as a catalyst for eliminating gender bias in AI systems, these ideas have not been empirically tested.

Our investigation has revealed that gender bias is hardly addressed in AI ethical guidelines issued by EU institutions or member states, despite all the ten analyzed documents proposing solutions to prevent discrimination. Future research should explore whether similar trends in addressing gender bias exist in other regions worldwide, which could inform best practices and activism in this domain.

Subsequent investigations should focus on quantitative assessments of the implementation of standards and tools to mitigate gender bias in AI systems by companies and organizations. Evaluations of solution effectiveness and potential conflicts between initiatives should also be considered. Such research has the potential to lay the foundation for the development of a grading system, quantifying the efforts of entities in preventing gender bias.

Appendix

Framework developed for the analysis of the guidelines. The framework consists of three sections: 1. Problem identification, 2. Action implementation and 3. Feedback.

Framework

Section 1: Problem identification

1.1. Does the document state where/how the AI technology is used within the company or organisation?

☐ Yes ☐ No

1.1a Do they develop the AI or integrate it in their product/service?

☐ Develop ☐ Integrate ☐ Other

1.1b Who is the product/service targeted at?

☐ Consumers ☐ Companies ☐ Other

1.2. Is gender biased identified and explained as an issue under the fairness and discrimination principle?

☐ Yes ☐ No

1.2a Does the issuer explain the potential cause/origin of gender discrimintaion within their company or organisation?

☐ Yes ☐ No

1.2a.1 Which are the causes explained?

☐ Consumers ☐ Companies

☐ Other

1.2b Does the document address all forms of discrimination simultaneously?

☐ Yes ☐ No

Section 2: Action Implementation

2.1. Does the issuer present/propose a potential solution(s) to prevent gender bias?

☐ Yes ☐ No

2.1a Are any specific solutions proposed?

☐ Checklist ☐ Training

☐ Algorithm ☐ Diversity team

☐ Other

2.1a.1 Is there a technical explanation (code) provided?

☐ Yes ☐ No

2.1b Will there be any consequences if the proposed solutiones are not followed/used?

☐ Yes ☐ No ☐ Not mentioned

2.2. Does the issuer present an implementation plan/approach for the proposed solutions?

☐ Yes ☐ No

2.2a Will everyone in the organisation or business be affected by this new methodology in the same way?

☐ Yes ☐ No

2.2b Are the proposed solutions meant to be used by individuals or by collectives?

☐ Individuals ☐ Collectives ☐ Both

Section 3: Feedback

3.1. Is there additional information/proof that the proposed solutions have been successfully/unsuccessfully implemented?

☐ Yes ☐ No

3.1a Has feedback been received from the different stakeholders involved in the process?

☐ Yes ☐ No

3.1a.1 Has this been positive or constructive feedback?

☐ Yes ☐ No

3.1b Has any literature been published on the proposed methodology?

☐ Yes ☐ No

3.1c Are other businesses or organisations using the same approach?

Funding Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Dove, G., Halskov, K., Forlizzi, J., Zimmerman, J.: UX Design Innovation. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 278–288. ACM, New York, NY, USA (2017)
2. Yang, Q.: Machine learning as a UX design material: how can we imagine beyond automation, recommenders, and reminders?. In AAAI Spring Symposia, Vol. 1, No. 2.1, pp. 2–6 (2018)
3. Figoli, F.A., Mattioli, F., Rampino, L.: Artificial Intelligence in the design process: the impact on creativity and team collaboration. FrancoAngeli, Milano, Italy (2022)
4. Antonelli, P.: AI Is Design's Latest Material. In: Google Design Library. AI Is Design's Latest Material (2018). Accessed 9 May 2023
5. Ropohl, G.: Philosophy of socio-technical systems. Soc. Philos. Technol. Quarter. Electron. J. **4**, 186–194 (1999). https://doi.org/10.5840/techne19994311
6. Bijker, W.E.: Do not despair: there is life after constructivism. Sci. Technol. Human Values **18**, 113–138 (1993)
7. Bijker, W.E.: Of bicycles, bakelites, and bulbs: toward a theory of sociotechnical change. MIT Press, Cambridge, MA (1997)
8. Jones, A.J.I., Artikis, A., Pitt, J.: The design of intelligent socio-technical systems. Artif. Intell. Rev.. Intell. Rev. **39**, 5–20 (2013). https://doi.org/10.1007/s10462-012-9387-2
9. Latour, B.: Where are the missing masses? The sociology of a few mundane artifacts. In: Shaping Technology/Building Society: Studies in Sociotechnical Change, pp. 225–228. MIT Press, Cambridge, MA (1992)
10. Callon, M.: Actor-network theory—the market test. Soc. Rev. **47**, 181–195 (1999). https://doi.org/10.1111/j.1467-954X.1999.tb03488.x
11. Johnson, D.G., Verdicchio, M.: Reframing AI Discourse. Minds Mach. (Dordr) **27**, 575–590 (2017). https://doi.org/10.1007/s11023-017-9417-6
12. Sciannamè, M.: Machine Learning (for) Design. Towards designerly ways to translate ML for design education. PhD Dissertation, Politecnico di Milano. (2023)
13. Moss-Racusin, C.A., Dovidio, J.F., Brescoll, V.L., et al.: Science faculty's subtle gender biases favor male students. Proc. Natl. Acad. Sci. **109**, 16474–16479 (2012). https://doi.org/10.1073/pnas.1211286109
14. Friedman, B., Nissenbaum, H.: Bias in computer systems. ACM Trans. Inf. Syst. **14**, 330–347 (1996). https://doi.org/10.1145/230538.230561
15. Leavy, S.: Gender bias in artificial intelligence: the need for diversity and gender theory in machine learning. In: Proceedings of the 1st International Workshop on Gender Equality in Software Engineering, pp. 14–16. ACM, New York, NY, USA (2018)
16. Nass, C., Moon, Y., & Green, N.: Are machines gender neutral? Gender-stereotypic responses to computers with voices. J. App. Soc. Psychol. **27**(10), 864–876 (1997)
17. Nass, C., Moon, Y.: Machines and mindlessness: social responses to computers. J. Soc. Issues **56**, 81–103 (2000). https://doi.org/10.1111/0022-4537.00153
18. Moussawi, S., Koufaris, M., Benbunan-Fich, R.: How perceptions of intelligence and anthropomorphism affect adoption of personal intelligent agents. Electron. Mark. **31**, 343–364 (2021). https://doi.org/10.1007/s12525-020-00411-w
19. Watkins, H., Pak, R.: Investigating user perceptions and stereotypic responses to gender and age of voice assistants. Proc Human Factors Ergon. Soc. Annu. Meet. **64**, 1800–1804 (2020). https://doi.org/10.1177/1071181320641434
20. Algorithm Watch. AI Ethics guidelines global inventory. https://inventory.algorithmwatch.org/database (2022). Accessed 9 May 2023
21. Bieling, T.: Design (&) Activism: Perspectives on Design as Activism and Activism as Design. Mimesis (2019)
22. Julier, G.: Design activism as a tool for creating new urban narratives. in Cipolla, C., & Peruccio, P. (eds.) Changing the Change: Design Visions, Proposals and Tools, Proceedings. pp. 813–822. Allemandi Conference Press. ISBN: 9788842216704 (2008)
23. Morshedzadeh, E., Dunkenberger, M.B., Nagle, L., et al.: Tapping into community expertise: stakeholder engagement in the design process. Policy Design and Practice **5**, 529–549 (2022). https://doi.org/10.1080/25741292.2022.2157130
24 Peters, B.: Policy Problems and Policy Design. Edward Elgar Publishing (2018)
25 Bohnet, I.: What Works: Gender Equality by Design. Harvard University Press (2016)
26. Shields, S.A.: Speaking from the Heart: Gender and the Social Meaning of Emotion. Cambridge University Press (2002)
27. Vinuesa, R., Azizpour, H., Leite, I., et al.: The role of artificial intelligence in achieving the sustainable development goals. Nat. Commun.Commun. **11**, 233 (2020). https://doi.org/10.1038/s41467-019-14108-y
28. Cowgill, B., Dell'Acqua, F., Deng, S., et al.: Biased programmers? Or biased data? A field experiment in operationalizing AI ethics. SSRN Electron. J. (2020). https://doi.org/10.2139/ssrn.3615404
29. Sampson, O.: A lovely day. Interactions **28**, 84–86 (2021). https://doi.org/10.1145/3439841
30 Lindgren, S., Holmström, J.: Social science perspective on artificial intelligence. J. Digit. Soc. Res. (2020). https://doi.org/10.33621/jdsr.v2i3.65
31. Mills, S.: Feminist Stylistic. Routledge (1995)
32. Butler, J.: Gender Trouble Feminism and the Subversion of Identity. Routledge (1990)
33. Fuss, D.: Essentially Speaking Feminism, Nature and Difference, 1st edn. Routledge (1990)
34. Barzilai, G., & Rampino, L. (2020). Just a natural talk? the rise of intelligent personal assistants and the (hidden) legacy of ubiquitous computing. In International Conference on Human-Computer Interaction, pp. 18–39. Cham: Springer International Publishing
35. Agudo, U., Liberal, K.G.: The Emperor's automagical suit: An experiment on bias and underperformance in image recognition AI. https://medium.com/bikolabs/the-emperors-automagical-suit-769079287f9f (2020). Accessed 9 May 2023

36. Agudo, U., Arrese, M., Liberal, K. G., & Matute, H.: Assessing Emotion and Sensitivity of AI Artwork. Front. Psychol. **13**:879088 (2022)
37. Karutis, K.: Exploring gender and compensation. In: InVision. https://www.invisionapp.com/inside-design/designer-compensation-gender/ (2016). Accessed 9 May 2023
38. Clement, J.: Global tech industry workforce diversity 2019, by gender. https://www.statista.com/statistics/784647/tech-industry-workforce-diversity-gender/ (2020)
39 Abercrombie, G., Cercas Curry, A., Pandya, M., Rieser, V.: Alexa, Google, Siri: What are Your Pronouns? Gender and Anthropomorphism in the Design and Perception of Conversational Assistants. In: Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing, pp. 24–33. Association for Computational Linguistics, Stroudsburg, PA, USA (2021)
40. Sava, J.A.: Distribution of tech jobs in Europe in 2020, by gender. In: Statista. https://www.statista.com/statistics/1222522/tech-position-gender-share-europe/ (2021). Accessed 9 May 2023
41. Fernández Álvarez L (2020) I Seminario - Ciberfeminismo: Cultura, mujeres y acción en red: ¿Puede ser una máquina quitanieves machista? Avaliable at: https://www.youtube.com/watch?v=gXoAE4Qiois
42 Forlano, L.: Posthumanism and design. She Ji J. Des. Econ. Innov. **3**, 16–29 (2017). https://doi.org/10.1016/j.sheji.2017.08.001
43. Bendick, M., Nunes, A.P.: Developing the research basis for controlling bias in hiring. J. Soc. Issues **68**, 238–262 (2012). https://doi.org/10.1111/j.1540-4560.2012.01747.x
44. Bendick, M., Jackson, C., Romero, J.: Employment discrimination against older workers: an experimental study of hiring practices. J. Aging Soc. Policy **8**, 25–46 (1996)
45. Bertrand, M., Mullainathan, S.: Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination, Cambridge, MA (2003)
46 Johnson, S., Hekman, D., Chan, E.: If there's only one woman in your candidate pool, there's statistically no chance she'll be hired. Harv. Bus. Rev.. Bus. Rev. **26**, 1–7 (2016)
47. Chamorro-Prezumic, T., Akhtar, R.: Should Companies Use AI to Assess Job Candidates? Harvard Business Review (2019)
48. Cowgill, B. (2018). Bias and productivity in humans and algorithms: Theory and evidence from resume screening. Columbia Business School, Columbia University, 29.
49. Feldman, M., Friedler, S.A., Moeller, J., et al.: Certifying and Removing Disparate Impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 259–268. ACM, New York, NY, USA (2015)
50. Raghavan, M., Barocas, S., Kleinberg, J., Levy, K.: Mitigating bias in algorithmic hiring. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 469–481. ACM, New York, NY, USA (2020)
51. Benjamins, R., Salazar García, I.: El mito del algoritmo. ANAYA Multimedia (2020)
52. Kurita, K., Vyas, N., Pareek, A., et al.: Measuring bias in contextualized word representations. In: Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pp. 166–172. Association for Computational Linguistics, Stroudsburg, PA, USA (2019)
53. Schwemmer, C., Knight, C., Bello-Pardo, E.D., et al.: Diagnosing gender bias in image recognition systems. Socius **6**, 237802312096717 (2020). https://doi.org/10.1177/2378023120967171
54. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2
55. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. Minds Mach. (Dordr) **30**, 99–120 (2020). https://doi.org/10.1007/s11023-020-09517-8
56. Webster, J., Watson, R.T.: Analyzing the past to prepare for the future writing a literature review. MIS Quart. **26**, xiii–xxiii (2002)
57 vom Brocke, J., Simons, A., Riemer, K., et al.: Standing on the shoulders of giants: challenges and recommendations of literature search in information systems research. Commun. Assoc. Inf. Syst.. Assoc. Inf. Syst. (2015). https://doi.org/10.17705/1CAIS.03709
58. Nadeem, A., Abedin, B., Marjanovic, O.: Gender Bias in AI: A review of contributing factors and mitigating strategies. ACIS 2020 Proceedings. (2020), https://aisel.aisnet.org/acis2020/27
59. Thinyane, M., Goldkind, L.: A multi-aspectual requirements analysis for artificial intelligence for well-being. In: 2020 IEEE First International Workshop on Requirements Engineering for Well-Being, Aging, and Health (REWBAH), pp. 11–18. IEEE (2020)
60. Zeng, Y., Lu, E., & Huangfu, C. Linking artificial intelligence principles. In the Proceedings of the AAAI Workshop on Artificial Intelligence Safety (AAAI-Safe AI 2019), 2019.
61 Fjeld, J., Achten, N., Hilligoss, H., et al.: Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI. SSRN Electron. J. (2020). https://doi.org/10.2139/ssrn.3518482
62. Whittlestone, J., Nyrup, R., Alexandrova, A., Cave, S.: The role and limits of principles in AI ethics. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 195–200. ACM, New York, NY, USA (2019)
63. Benjamins, R.: Towards organisational guidelines for the responsible use of AI. In: 24th European Conference on Artificial Intelligence. ECAI 2020. (2020) arXiv:2001.09758
64. Dignum, V.: Ethics in artificial intelligence: introduction to the special issue. Ethics Inf. Technol. **20**, 1–3 (2018). https://doi.org/10.1007/s10676-018-9450-z
65. Wagner, B.: Ethics as an escape from regulation. From "ethics-washing" to ethics-shopping? In: BEING PROFILED, pp. 84–89. Amsterdam University Press (2019)
66. Benjamins, R., Barbado, A., Sierra, D.: Responsible AI by design in practice. In: Human-Centered AI: Trustworthiness of AI Models & Data (HAI). https://arxiv.org/abs/1909.12838 (2019)
67 Vermaas, P.E., Pesch, U.: Revisiting Rittel and Webber's Dilemmas: designerly thinking against the background of new societal distrust. She Ji J. Des. Econ. Innov. **6**, 530–545 (2020). https://doi.org/10.1016/j.sheji.2020.11.001
68. Utrecht Data School. Data Ethics Decision Aid (DEDA). https://dataschool.nl/en/deda/ (2021). Accessed 9 May 2023
69. Club Informatique des Grandes Entreprises Françaises (CIGREF). Digital ethics: a guide for professionals of the digital age. https://www.cigref.fr/digital-ethics-guide-professionals-the-digital-age-cigref-syntec-2018 (2018)
70. Council of Europe. Artificial intelligence and data protection. https://edoc.coe.int/en/artificial-intelligence/8254-artificial-intelligence-and-data-protection.html (2019)
71. Rasmussen, S., & The Expert Group on November 2018 DATA ETHICS. Data for the Benefit of the People: Recommendations from the Danish Expert Group on Data Ethics. Økonomi- og Erhvervsministeriet. https://em.dk/media/12190/dataethics-v2.pdf (2018)

72. SAP. SAP's Guiding Principles for Artificial Intelligence. https://www.sap.com/documents/2018/09/940c6047-1c7d-0010-87a3-c30de2ffd8ff.html (2021)
73. Independent High-Level Expert Group on Artificial Intelligence. Ethics Guidelines For Trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (2019)
74. Telefónica, S.A.: Enfoque de Telefónica para un uso responsable de la IA. https://www.telefonica.com/es/wp-content/uploads/sites/4/2021/08/ia-uso-responsable.pdf (2018)
75. Eswaran, V.: The business case for diversity in the workplace is now overwhelming. https://www.weforum.org/agenda/2019/04/business-case-for-diversity-in-the-workplace/ (2019). Accessed 9 May 2023
76. Villani, C.: For a meaningful artificial intelligence. https://www.jaist.ac.jp/~bao/AI/OtherAIstrategies/MissionVillani_Report_ENG-VF.pdf (2018)
77. Pennington, C.: We are hardwired to resist change. In: Emerson Human Capital. https://www.emersonhc.com/change-management/people-hard-wired-resist-change (2018). Accessed 9 May 2023