AI & SOCIETY
https://doi.org/10.1007/s00146-024-02067-y

OPEN FORUM

Mitigation measures for addressing gender bias in artificial intelligence within healthcare settings: a critical area of sociological inquiry

Anna Isaksson[1]

Received: 7 May 2024 / Accepted: 4 September 2024

Abstract
Artificial intelligence (AI) is often described as crucial for making healthcare safer and more efficient. However, some studies point in the opposite direction, demonstrating how biases in AI cause inequalities and discrimination. As a result, a growing body of research suggests mitigation measures to avoid gender bias. Typically, mitigation measures address various stakeholders such as the industry, academia, and policy-makers. To the author's knowledge, these have not undergone sociological analysis. The article fills this gap and explores five examples of mitigation measures designed to counteract gender bias in AI within the healthcare sector. The rapid development of AI in healthcare plays a crucial role globally and must refrain from creating or reinforcing inequality and discrimination. In this effort, mitigation measures to avoid gender bias in AI in healthcare are central tools and, therefore, essential to explore from a social science perspective, including sociology. Sociologists have made valuable contributions to studying inequalities and disparities in AI. However, research has pointed out that more engagement is needed, specifically regarding bias in AI. While acknowledging the importance of these measures, the article suggests that they lack accountable agents for implementation and overlook potential implementation barriers such as resistance, power relations, and knowledge hierarchies. Recognizing the conditions where the mitigation measures are to be implemented is essential for understanding the potential challenges that may arise. Consequently, more studies are needed to explore the practical implementation of mitigation measures from a social science perspective and a systematic review of mitigation measures.

Keywords Artificial intelligence · Gender bias · Mitigation measures · Healthcare sector

1 Introduction

Artificial intelligence (AI) is rapidly transforming medicine and the healthcare sector. It is often described as crucial in making future healthcare safer and more efficient. Integrating AI in healthcare is considered necessary for addressing challenges arising from an aging population and the shortage of proficient healthcare staff (Bajwa et al. 2021). AI can provide treatment recommendations and prioritization decisions when resources are limited (Alowais et al. 2023).

Previous studies have reported how AI has been developed and used in various areas, such as identifying embryos in in-vitro fertilization, assessing X-ray images, and detecting various forms of cancer (Esteva et al. 2017; Khosravi et al. 2019; Liu et al. 2019; Hu et al. 2019). For computer-aided diagnosis and image-based screening, AI systems are being adopted worldwide (Larrazabal 2020). Moreover, studies show how AI can predict risks of conditions such as sepsis and cardiovascular disease (Goh et al. 2021; Oikonomou et al. 2019; Weng et al. 2017).

However, while the use of AI in healthcare is expected to contribute to high-quality care, some studies point in the opposite direction. AI can reinforce biases, inequality, and discrimination. Understanding AI technologies' impacts is thus significant (O'Connor and Liu 2023; Marinucci et al. 2023).

In their systematic literature review, Lau et al. (2023) conclude that there is a need for more representative data on women's healthcare conditions on a general level. Similarly, García-Micó and Laukyte (2023) highlight how female

✉ Anna Isaksson
anna.isaksson@hh.se

[1] School of Health and Welfare, Halmstad University, Halmstad, Sweden

Published online: 18 September 2024

representation and bodies are underrepresented in training algorithms. According to the authors, a massive data gap exists between females' and males' involvement in clinical trials, medical therapies, and disease treatment. Representative datasets are important since diseases may manifest differently in women and men, younger and older individuals, and across different ethnicities. Consequently, unrepresentative training data used by algorithms in healthcare can lead to misdiagnoses and biased, ineffective, and discriminatory AI applications. In healthcare, such discrimination can affect which groups receive accurate diagnoses and appropriate treatments. Further, machine-learning models learn from historically collected data that could be based on stereotypes. When operationalized into decision-making systems in healthcare, this can have significant consequences (c.f. Rajkomar et al. 2018).

As a result of the growing body of research indicating significant risks of bias in AI, there has been a general increase in research addressing how risks can be mitigated (e.g. Bellamy et al. 2019; Deshpande et al. 2020; Dhar et al. 2020; Stafanovičs et al. 2020; Wang et al. 2019). Mitigation measures, checklists, and recommendations to avoid risks have been developed and presented in research articles, policy documents, research funding bodies, and journal instructions for authors who report their findings within AI research. Some mitigation measures are targeted directly at medicine and healthcare, while others are broader and refer to AI research in various fields. Some focus only on gender, while others concern social categories such as age, disability, and ethnicity. However, to the author's knowledge, these kinds of mitigation measures have not yet undergone sociological analysis.

According to Liu (2021), there is an increasing interest in utilizing sociological concepts to examine AI's social aspects, origins, and outcomes. Sartori and Theodorou (2022) also discuss how a sociological approach to AI expands the exploration of its broad societal implications and provides essential insights for AI developers and designers. Current research is driven by technical possibilities applied to social and economic issues. This technical focus often overlooks social complexities like inequality, diversity, and power dynamics. In contrast, sociologically informed research questions could significantly influence and enhance technological development, ensuring AI systems are more attuned to these social nuances. Further, Joyce et al. (2021) argue that sociologists can identify how inequalities are embedded in all aspects of society and, therefore, have a prominent role in understanding the development of AI since AI is not just a question of technological design and development. The development and implementation of AI raises questions about power and social order. Sociologists have made valuable contributions to studying inequalities and disparities in AI design, development, and use. However, more engagement from sociological scholars is needed, specifically regarding issues such as bias in AI (Joyce et al. 2021; Zajko 2022), which this article targets.

Consequently, this article aims to explore and analyze five examples of mitigation measures proposed in research within medicine and healthcare. In doing so, the article contributes to the sociology of AI – an emerging research field that, according to Liu (2021), engages scholars from across the full spectrum of sociology. Further, the article raises practical implications for organizations that make use of mitigation measures in practice.

Two research questions are posed in the article:

1. Which pathways toward mitigating bias are suggested?
2. What actors do the mitigation measures target?

The article analyzes mitigation measures and recommendations developed to counteract bias in AI, specifically within the healthcare sector. The development of this sector plays a crucial role globally. Today, it is evident that AI affects or will affect almost every aspect of healthcare (Chen and Decary 2020). Hence, exploring mitigation measures from several perspectives within social sciences, including sociology, is important.

Mitigation measures are understood as specific practices, actions, recommendations and/or strategies taken to reduce and/or prevent undesirable impacts of AI while still allowing and supporting it (cf. Straw and Stanley 2023). In a healthcare context, mitigation measures are utilized for reducing bias in the design, development and implementation of AI systems. They aim to minimize risks and maximize the benefits of medical AI (c.f. Gray et al. 2024: European Parliament 2022).

The article is structured as follows. First, the methodological framework of the article is presented. Next, the article outlines examples of the implications of gender bias in AI within healthcare settings, aiming to contextualize the background and imperative for mitigation measures. The following sections present and analyze five cases of mitigation measures for avoiding AI bias. The article concludes with a discussion that reconnects to the purpose of the article. Finally, this article suggests lessons that can be learned from these cases going forward.

2 Methodological framework

This study applies a case study methodology, an approach used in previous studies interested in gender bias perpetuation and mitigation in AI technologies (O'Connor and Liu 2023). Initially, searches were conducted in databases (Scopus, PubMed, and Cinahl) based on various combinations of keywords such as "Gender bias," "Artificial intelligence,"

"Healthcare," "Mitigation measures," "Recommendations," and "Avoiding bias in AI." The five case studies that became the subject of analysis in this study were chosen because they met the following criteria: (a) Consist of recommendations for avoiding bias in AI in the healthcare sector, (b) Emphasize social rather than technological aspects in the mitigation measures, and (c) Newly published (2020–2023) articles in peer-reviewed journals. The selection of the examples aligns with the article's ambition to present and analyze examples of mitigation measures and to draw attention to these measures as a critical area of sociological inquiry and calling for further exploration from a social science perspective.

A case study is typically utilized when the researcher is interested in comprehensively understanding a specific individual or group and/or, like in this article, a specific phenomenon. It often employs various methods like interviews, observation, and text analysis. The disadvantage usually emphasized is that case studies cannot be generalized. On the other hand, some argue that case studies are generalizable when developing and generalizing theories. However, a case study cannot generalize populations or describe frequencies (Yin 2018; Merriam 2009). This article does not intend to generalize. Instead, nuances and complexities are explored to gain insights into the specific phenomenon of mitigation bias to avoid gender bias in AI in healthcare settings.

3 Gender bias in AI in healthcare settings – the imperative for mitigation measures

Today, there is no generally accepted definition of artificial intelligence (Sheikh et al. 2023). According to Bajwa et al. (2021: 189), AI refers to:

> (…) the science and engineering of making intelligent machines through algorithms or a set of rules, which the machine follows to mimic human cognitive functions, such as learning and problem-solving. AI systems have the potential to anticipate problems or deal with issues as they come up and, as such, operate in an intentional, intelligent, and adaptive manner.

A general and simplified understanding is that AI can learn and recognize patterns from large datasets. This means that AI systems in healthcare settings can, for instance, identify various diseases and provide treatment recommendations and prioritization decisions. AI can also convert medical records into information indicating likely diagnoses (Quinn et al. 2021).

Among the most common causes of AI bias reproduced in a healthcare setting are imbalanced datasets within the AI training data. Larrazabal et al. (2020) discuss how AI systems trained primarily on males perform worse when tested on females. This can be derived from diseases manifesting differently for women and men. Taking dermatology as an example, sex and gender differences are essential to consider when creating algorithms due to gender-based differences in skin diseases that can affect the performance of AI in skin cancer diagnosis. For example, females suffer more frequently from melanomas on the hip and lower extremities compared to males. Therefore, algorithms must rely on training sets with a variety and diversity of patients, taking different dermatologic conditions for different sexes into account (Lee et al. 2022; Olsen et al. 2020; Yuan et al. 2019).

Similar examples have been identified in other areas. In their study, Tomašev et al. (2019) as cited in Ibrahim et al. 2021) examine an algorithm for predicting acute kidney injury in adults. The training dataset was unrepresentative and derived mainly from male patients. As expected, the algorithm was found to cause gender disparities and bias. However, previous research has also demonstrated strategies to overcome challenges with unrepresentative datasets. Such research has shown that it is possible to be gender sensitive in the development of AI. Cardiovascular disease has, for example, different expressions in women and men, which has been recognized decades ago. Symptoms of cardiovascular disease are more subtle in women, resulting in delayed diagnosis and treatment (Baart et al. 2019). Consequently, it is crucial to consider gender when building a cardiovascular disease diagnosis model. In a recent study, the effect of separating the dataset of cardiovascular disease patients into two different datasets for female and male patients was investigated. Separate dataset diagnosis models for females and males resulted in faster detection, decision-making, and treatment (Hogo 2020).

Generally, gender bias can arise when algorithms are trained using unrepresentative datasets. However, existing biases, norms, and stereotypes in society and healthcare can also influence AI bias's (re)production. As argued by Hassani (2021), the way society behaves (for example, discrimination towards gender) will be reflected by the system and models since algorithms learn from the data provided. If data contains gender stereotypes and are affected by a social bias perspective, the AI system will perpetuate these disparities (Saka 2020; Rajkomar et al. 2018).

Previous studies have highlighted that a lack of diversity and interdisciplinarity in AI development teams could also cause AI biases and discrimination. It has been argued that including women in the development and implementation of AI is crucial since AI solutions tend to be designed with limited (male) perspectives and experiences. This affects the quality of the solutions (Roopaei et al. 2021). However, female representation in AI is progressing slowly. Still, developers of AI are overwhelmingly male. UNESCO reports that women represent only 12% of artificial intelligence researchers globally, 20% of employees in technical

roles in machine learning companies, and 6% of professional software developers (World Economic Forum 2023).

In sum, previous research has identified several reasons why gender bias in AI arises. When missing data and/or data based on bias/stereotypes is operationalized into algorithmic decision-making systems, discrimination and gender disparities will occur. This poses particular risks in sectors like healthcare (Larrazabal et al. 2020). Therefore, a growing body of studies has recognized an urgent need for mitigation measures, recommendations, and/or checklists to minimize the risks associated with AI applications and maximize the benefits of AI in healthcare contexts.

4 Cases of mitigation measures to avoid gender bias in AI

In this section, five examples (Buslón et al. 2023; Cirillo et al. 2020; Wesson et al. 2022; Gichoya et al. 2023; Lee et al. 2022) of mitigation measures to avoid gender bias in AI in healthcare are presented.

In their article, Buslón et al. (2023: 01) present "(…) key recommendations to facilitate the inclusion of sex and gender perspective into public policies, educational programs, industry, and biomedical research, among other sectors, and help overcome sex and gender biases in AI and health". The recommendations result from awareness-raising actions such as keynote sessions, seminars, and working groups with invited academic experts, journalists, NGOs, industry, and policy-makers. Buslón et al. (2023:05–06) present recommendations to various target groups such as AI practitioners and academies, industry, civil society, and policy-makers and governments. The recommendations for each target group are summarized in the table below:

Target group	Recommendation 1	Recommendation 2	Recommendation 3	Recommendation 4	Recommendation 5
AI practitioners and academies	Educate and train on how to introduce the sex and gender dimension in research and during career-specific subjects focusing on the social impact of AI and its applications for social good	Define and apply metrics to evaluate AI biases, and techniques for mitigating biases in datasets and models	Use adequate guidelines and international recommendations to standardize the data collection process	Work and include a socially inclusive and interdisciplinary perspective to define cross-cutting solutions in health and AI that include health professionals, sociologists, psychologists, engineers and end-users to empower diverse teams with better results and solutions	Promote publishing policies that foster sex and gender balance in research recruitment and analysis of data
Industry	Promote the use of certification and regulation in all the processes	Provide more information to all the staff regarding the relevance of quality and fairness in training data sets	Develop and apply methods to balance databases and make more emphasis on the documentation of AI processes and applicability	Include socio-cultural aspects into AI processes by integrating different points of view in the design, development and evaluation of technological research in order to avoid bias and to generate audits for citizen participation	
Civil society	Participate in educational programs, open debates, and other initiatives to grow a digital literacy and responsible scientific research and development culture in a plural and egalitarian way	Require scientific evidence and transparency in AI processes from companies, academia, and policy-makers that have a social impact	Develop initiatives that protect the most vulnerable population, especially women, in AI applications and health	Promote and demand AI guidelines aligned with the UN SDGs as a goal to have clear benefits in society	

Target group	Recommendation 1	Recommendation 2	Recommendation 3	Recommendation 4	Recommendation 5
Policy-makers and governments	Define and offer a public certification and regulation as a key legal aspect in all sectors to guarantee the benefits of AI in society	Provide an inclusive strategy in health promotion and care to benefit all citizens	Develop public policies in trustworthy AI to apply in all sectors	Invest in research and initiatives on AI regarding gender and diversity	Provide more public information to citizens and training about how to introduce the perspective of sex and gender in science

Cirillo et al. (2020) explore sex and gender disparities in biomedical technologies used in precision medicine. They argue that biomedical AI tools lack bias detection mechanisms and risk overlooking the influence of sex and gender on health and disease. Based on their study, Cirillo et al. provide four recommendations "(…) to ensure that sex and gender differences in health and disease are accounted for in AI implementations that inform Precision Medicine (2020: 81)". The integration of these four recommendations aims to accelerate progress toward developing effective strategies to enhance population health and wellbeing.

> Distinguish between desirable and undesirable biases and guarantee the representation of desirable biases in AI development (see Introduction: Desirable vs. Undesirable biases).
> Increase awareness of unintended biases in the scientific community, technology industry, among policy-makers, and the general public (see Sources and types of Health data and Technologies for the analysis and deployment of Health data).
> Implement explainable algorithms, which not only provide understandable explanations for the layperson, but which could also be equipped with integrated bias detection systems and mitigation strategies, and validated with appropriate benchmarking (see Valuable outputs of Health technologies).
> Incorporate key ethical considerations during every stage of technological development, ensuring that the systems maximize wellbeing and health of the population (…).

Gichoya et al. (2023) discuss how diverse AI applications are utilized across various healthcare systems. As their usage expands, failures of these applications and how they can perpetuate bias are discovered. Therefore, according to the authors, it is crucial to prioritize bias assessment and mitigation, particularly in radiology applications. They present pitfalls within the larger AI lifecycle—from problem definition and dataset selection to model training and deployment—causing AI bias. Based on these pitfalls, strategies for mitigating biases within the broader framework of implementing AI in the healthcare enterprise are provided.

They summarize their recommendations as follows:

To mitigate bias, we must continue to create and use diverse and representative data sets develop and test rigorous testing and validation protocols, perform ongoing monitoring and evaluation of model performance. Bias risk assessment tools can facilitate this process. Bias mitigation is especially important as we understand human–machine partnership, with early findings showing worsening performance for experts when presented with biased model outputs. Most importantly, we cannot overemphasize the need for diverse teams to work on this challenging topic. By taking a comprehensive and multifaceted approach to addressing bias in AI model development, researchers and practitioners can help to ensure that these technologies are used ethically and responsibly to benefit all patients (Gichoya et al. 2023: 6).

Lee et al. (2022) discuss the growing interest in utilizing machine learning and artificial intelligence tools in dermatology. They are particularly interested in skin cancer diagnosis and evaluation of other dermatologic conditions. As these technologies evolve, it is critical to mitigate disparities based on sex and gender in healthcare delivery. Lee et al. advocate for the consideration of sex and gender differences in algorithms within dermatology due to sex-specific conditions of various cancers and autoimmune disorders. They propose the following recommendations aiming to enhance favoring bias while preventing undesirable bias:

- Incorporate sex and gender into patient metadata when creating algorithms
- Report patient demographics of training and testing datasets
- Demonstrate fair performance across the spectrum of sexes and genders
- Account for intersectionality of patient factors, such as race, sexuality, and gender, in algorithms for increased accuracy
- Ensure adequate representation of gender and racial minorities
- Demonstrate accuracy of performance for specific, marginalized groups
- Increase reporting of sex and gender distributions for ML/AI datasets

- Ensure adequate representation of all sexes and genders in datasets
- Promote future research on differences in dermatologic conditions in all genders, including transgender, nonbinary, and other gender-diverse patients (Lee et al. 2022:402).

In their review, Wesson et al. (2022) discuss big data and health equity concerns. They explore instances where big data applications unintentionally have reinforced discriminatory practices. Big data is contextualized through the lens of the five Vs—(volume, velocity, veracity, variety, and value), and a sixth V is introduced – virtuosity. Virtuosity targets equity and justice frameworks. Wesson et al. illustrate analytical approaches for enhancing equity and suggest that the following recommendations should be taken into account to improve equity in big data research:

- First, include social epidemiologists in research and prioritize social epidemiology training beyond programs in epidemiology and other public health disciplines. Scientists who study social epidemiology have deeper knowledge of the structural and systemic forces that have generated the distribution of advantages and disadvantages in society (…)
- Second, increase the level of diversity in researchers across disciplines pursuing big data and equity. Discriminatory biases can be prevented through the addition of a wide range of perspectives, which can reduce the likelihood of generating biases based on singular viewpoints (…)
- Third, generate partnerships between industry and academia (…). Big tech should work with social epidemiologists to generate more ethical and virtuous research.
- Fourth, federal and state policies are needed to safeguard against biased and discriminatory production of big data.
- Fifth, it is important for us as scientists to evaluate our own biases and understand that we do not have the breadth of experience to know what is fully needed to improve equity. (Wesson et al. 2022: 70–71).

5 Analysis and discussion

The five cases of mitigation measures presented provide comprehensive recommendations, actions, and strategies for preventing AI bias in healthcare and medicine. Even if the mitigation measures concern different areas and aspects of healthcare, the following analysis reveal how they target some similar issues. This analysis and discussion section is organized into three subthemes that identify, examine, and critically discuss recurrent patterns observed across the five cases.

5.1 Inclusivity and diversity as mitigation pathways

The importance of inclusivity and diversity is a prominent theme in the five cases, meaning that interdisciplinarity, interdisciplinary collaboration, and diverse teams are emphasized in various ways. Gender and diversity perspectives are distinguished as crucial for creating AI technologies and systems that are fair and unbiased. Buslón et al. (2023) recommend, for example, that a socially inclusive and interdisciplinary perspective should be included and that results and solutions will be better with diverse teams that include health professionals, sociologists, psychologists, engineers, and end-users. Promoting publishing policies that foster sex and gender balance in recruitment and analysis of data are also highlighted in their recommendations. Further, they also recommend that socio-cultural insights into AI processes should be included by integrating different points of view in the research's design, development, and evaluation. Gichoya et al. (2023) argue that they cannot overemphasize the need for diverse teams to work on this challenging topic. They recommend a comprehensive and multifaceted approach to addressing bias in AI. Inclusive datasets when creating algorithms by incorporating sex and gender, racial minorities, transgender, nonbinary, and other gender-diverse patients are also highlighted in Lee et al. (2022) while Wesson et al. (2022) recommend diversity in researchers across disciplines. Further, they argue that increasing the level of diversity in researchers can prevent discriminatory biases through a wide range of perspectives. Generating partnerships between industry and academia is also recommended.

In summary, embracing inclusivity and diversity, emphasizing collaboration between practioners and academia, acknowledging interdisciplinarity, and indirectly challenging the notion that one scientific discipline is superior to another in the context of medical AI is promising. Much evidence supports an urgent need for engineers and researchers in gender studies, social sciences, design, technology, data science, et cetera, to collaborate in developing fair AI. As Schiebinger and Schraudner (2011) argue, gender analysis and perspective can provide critical rigour in medicine research, policy and practice. However, the recommendations give the impression that there are no knowledge hierarchies and conflicts in working in an interdisciplinary manner and adopting a gender perspective. The potential lack of understanding and communication between technology-oriented and social science perspectives is overlooked, as well as the fact that researchers and practitioners in these areas may have different terminologies, methods, and goals. This can make it challenging to work with these perspectives in an effective and integrated manner. Although there seems to be agreement behind the recommendations that interdisciplinarity (or transdisciplinarity) knowledge production can overcome bias in AI, it is likely that this knowledge production heavily

depends on broader changes to the current knowledge regime in place (c.f. Felt et al. 2016).

The incorporation and consideration of gender and other social categories are depicted as processes that can be seamlessly undertaken. Even if some argue that there has been much less resistance in the research community against gender issues and perspectives during the last decades (Mellström, 2021), it can also be noted that we are facing a new wave of resistance to gender theory, perspectives and practice (Kuhar and Paternotte 2017). The recommendations are expected to be implemented in contexts likely characterized by certain conditions, power relations, and knowledge hierarchies. While significant progress has been achieved through interdisciplinary teams and integrating gender and diversity perspectives into health-related research, there are still challenges. It cannot be taken for granted that researchers are prepared to conduct and/or accept sophisticated sex and gender analysis (c.f. Schiebinger and Klinge 2015).

5.2 Enhancing knowledge and awareness raising activities

In addition to recommendations highlighting the importance of inclusivity and diversity, increasing knowledge and raising awareness about bias in AI are emphasized in various ways. To exemplify, Wesson et al. (2022) call for reflexivity among scientists and suggest that scientists should evaluate their own biases and understand that they do not know what is needed to improve equity. Cirillo et al. (2020) call for increased awareness of unintended biases among the general public, policy-makers, scientific community stakeholders, and the technology industry. Further, Buslón et al. (2023) address various stakeholders in their recommendations. AI practitioners and academics are recommended to offer information to all employees about the importance of quality and fairness in training datasets. The authors recommend that AI practitioners and academies provide education and training on incorporating the sex and gender dimension into research and career-specific subjects, emphasizing the social impact of AI. They suggest the industry engages in educational programs, open discussions, and additional initiatives to foster digital literacy and cultivate a diverse culture of responsible scientific research and development. Policy-makers and governments should increase public awareness by offering citizens more information and training on incorporating sex and gender perspectives into science.

In summary, some of the recommendations in the five cases underline the significance of enhancing knowledge and fostering awareness regarding sex, gender, and unintended bias. Given the consequences of overlooking gender and diversity perspectives in the development and utilization of AI technology in healthcare, it becomes apparent that there is a need for more gender-sensitive knowledge. However, it is crucial to remind oneself that these recommendations should be implemented in a masculine-coded sector, as men generally dominate several technical professions related to AI (World Economic Forum, 2023). There is still a strong relationship between technology and masculinity—in other words, technology and the masculinization of power are intimately connected (cf. Holth and Mellström 2011). Furthermore, it can be noted that this type of recommendation assumes that increased knowledge automatically leads to action/change. There is far from a causal relationship between increased knowledge and action/change regarding gender and diversity issues. Previous research has shown how resistance is activated when gender-based knowledge is to be integrated into an organization and/or research. Research conducted in hierarchical organizations with gendered power relations, such as academia, has demonstrated that organizational receptivity is needed to translate gender-based knowledge into action. It requires that the knowledge is seen as legitimate (Jordansson and Peterson 2024; Schiebinger and Schraudner 2011).

5.3 Abscense of responsibility and accountability

Overall, there is an absence of agents and implementation responsibility in the mitigation measures explored. Indeed, Buslón et al. (2023:05–06) present recommendations for different stakeholders and sectors, but on a very overarching level. For all mitigation measures, it is notable that the organizations and stakeholders who are potential receivers of the recommendations do not consist of actual people and functions. For instance, Cirillo et al. (2020) recommend incorporating key ethical considerations during every technological development stage. However, the entity or function responsible for "incorporating" is not specified. Buslón et al. (2023:05–06) recommend AI practitioners and academies to "(…) Use adequate guidelines and international recommendations to standardize the data collection process". Here, it is unclear who should ensure this is done and followed up. Nor is it clear who or what determines "adequate guidelines."

Another example illustrating the overall absence of actors and subjects responsible for action and implementation is a recommendation by Lee et al. (2022). The authors suggest that future research on differences in dermatologic conditions in all genders, including transgender, nonbinary, and other gender-diverse patients, should be promoted. However, it is unclear who is responsible for promoting research and whether this advice should be directed towards individuals or groups and/or at an institutional or structural level.

Indeed, it is challenging to pinpoint who should be held accountable for implementing the mitigation measures. The recommendations have broad application areas, and the organizations in which they are to be implemented can vary in structure and consist of different roles and functions.

Therefore, it is reasonable to consider that actors and subjects who can be held accountable cannot be precisely identified. However, the point is that the recommendations are received in local organizations and settings where "translation work" must be done regarding who is responsible for what actions. In previous research, it has been emphasized that senior management typically must be held accountable for the proper implementation of gender analysis. The importance of gender-sensitive leadership that authorizes the implementation of gender-based knowledge should not be underestimated (Jordansson and Peterson 2024; Schiebinger and Schraudner 2011). Further, a textually sanctioned agency produces a power that potentially mobilizes people's actions and works. Similarly, history has shown us that when organizational categories, roles, and functions are rendered invisible, and it is unclear whose responsibility it is to ensure the integration and implementation of gender issues, it can result in inaction (cf. Powell 2018).

6 Conclusions and recommendations for future research

The aim of this article has been to explore and analyze five examples of mitigation measures proposed in research within medicine and healthcare. Two research questions were addressed:

Which pathways toward mitigating bias are suggested?
What actors do the mitigation measures target?

In previous section, two prominent pathways for mitigating bias were identified and explored. The first pathway circled around inclusivity and diversity, including interdisciplinary collaboration, diverse teams, and integrating gender perspectives in AI development. These recommendations assume no knowledge hierarchies or potential conflicts in interdisciplinary work, overlooking differences in terminology, methods, and goals between social science and technology perspectives. The second pathway emphasized the importance of enhancing knowledge and awareness raising activities. This pathway included reflexivity among scientists and the promotion of educational activities. These recommendations are based on underlying assumptions that there is a causal relationship between increased knowledge and action, neglecting potential conflicts of interest and various forms of resistance.

The analysis related to the second research question highlighted that accountable actors are absent overall in the mitigation measures. A need for more clarity concerning implementation responsibility in the recommendations was suggested, and it was argued that the absence of agents may hinder the "translation of recommendations" into action within local organizations and settings.

In conclusion, the mitigation measures analyzed focus on how to counteract AI bias in healthcare settings, thereby addressing societal inequality and discrimination—social issues at the heart of sociological research. Paradoxically, these recommendations are formulated in a way that overlook the inequalities that may characterize the context in which they are intended to be implemented. Hence, this article emphasizes the importance of recognizing the conditions where mitigation measures are to be implemented to understand potential challenges. Implementing these recommendations effectively requires that possible power relations and knowledge hierarchies in targeted organizations are recognized and there is an organizational receptivity and acceptance. These critical insights are essential for organizations to translate recommendations into practice successfully. The article also has important policy implications. Health equity is a primary goal of national and international institutions and various healthcare stakeholders. At the EU level, for example, there are comprehensive policy recommendations aimed at minimizing and mitigating the risks of AI in healthcare and strengthening its ethical and responsible development. (European Parliament 2022). This article suggests that policymakers, with support from researchers, need to evaluate their policy efforts to ensure it is having the intended effect.

Finally, this article should not be interpreted as a critique of the mitigation measures analyzed. Instead, the findings remind us that sociology and other areas within social science provide valuable insights when exploring the contexts in which mitigation measures are to be implemented. Contexts that, with support from previous research, are likely characterized by knowledge hierarchies and power relations. As mentioned, sociology has raised concerns about inequalities and disparities in AI design, development, and use. Following Liu (2021) and Joyce et al. (2021), who stress that more engagement from sociological scholars is needed, this article has contributed by identifying a new critical area of sociological inquiry—mitigation measures for avoiding gender bias in AI. Based on the results in this article, examples of areas for future research include a more systematic review of mitigation measures for avoiding AI bias in healthcare. A study of this nature could provide a more thorough exploration of the characteristics of mitigation measures than the current study has been able to accomplish. This article has also revealed that there is a need to investigate policy efforts and the implementation of mitigation measures further. Therefore, sociology—focusing on structural change and inequalities—has additional contributions to make, as well as other disciplines within social science targeting organizational issues, knowledge regimes and hierarchies, leadership and organization, and cultural change. Possible

research questions may be: Under what conditions are the mitigation measures likely to be accepted and implemented? How can potential resistance related to knowledge hierarchies and power relations be prevented? What organizational receptiveness is needed to implement mitigation measures successfully?

Funding Open access funding provided by Halmstad University. Åke Wiberg Stiftelse,H22-0022,Anna Isaksson

Data availability The data that support the findings of this study are openly available (Open access): https://doi.org/https://doi.org/10.1038/s41746-020-0288-5. https://doi.org/https://doi.org/10.1259/bjr.20230023. https://doi.org/https://doi.org/10.1093/jamia/ocab113. https://doi.org/https://doi.org/10.3389/fgwh.2023.970312. https://doi.org/https://doi.org/10.1146/annurev-publhealth-051920-110928.

Declarations

Conflict of interest The corresponding author states that there is no conflict of interest.

Ethical approval An ethics statement and informed consent are not applicable because this study is based exclusively on published literature.

References

Alowais SA, Alghamdi SS, Alsuhebany N et al (2023) Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Education. https://doi.org/10.1186/s12909-023-04698-z

Baart SJ, Dam V, Scheres LJJ, Damen JAAG et al (2019) Cardiovascular risk prediction models for women in the general population: a systematic review. PLoS One. https://doi.org/10.1371/journal.pone.0210329

Bajwa J, Munir U, Nori A et al (2021) Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthcare J 8(2):188–194. https://doi.org/10.7861/fhj.2021-0095

Bellamy RK, Dey K, Hind M et al (2019) AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. IBM J Res Dev 63(4/5):1–4

Buslón N, Cortés A, Catuara-Solarz S et al (2023) Raising awareness of sex and gender bias in artificial intelligence and health. Front Global Women's Health. https://doi.org/10.3389/fgwh.2023.970312

Chen M, Decary M (2020) Artificial intelligence in healthcare: an essential guide for health leaders. Healthc Manage Forum 33(1):10–18. https://doi.org/10.1177/0840470419873123

Cirillo D, Catuara-Solarz S, Morey C et al (2020) Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. npj Digital Medicine. https://doi.org/10.1038/s41746-020-0288-5

Deshpande KV, Pan S, Foulds JR (2020). Mitigating demographic Bias in AI-based resume fltering. In Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization (pp. 268–275).

Dhar P, Gleason J, Souri H, Castillo CD, Chellappa R (2020) Towards gender-neutral face descriptors for mitigating bias in face recognition. arXiv preprint arXiv:2006.07845

Esteva A, Kuprel B, Novoa R et al (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115–118. https://doi.org/10.1038/nature21056

Felt U, Igelsböck J, Schikowitz A, Völker T (2016) Transdisciplinary sustainability research in practice: between imaginaries of collective experimentation and entrenched academic value orders. Sci Technol Human Values 41(4):732–761. https://doi.org/10.1177/0162243915626989

García-Micó TG, Laukyte M (2023) Gender, Health, and AI: How Using AI to Empower Women Could Positively Impact the Sustainable Development Goals. In: Mazzi F, Floridi L (eds) The Ethics of Artificial Intelligence for the Sustainable Development Goals. Philosophical Studies Series, Springer, Cham

Gichoya JW, Thomas K, Celi LA et al (2023) AI pitfalls and what not to do: mitigating bias in AI. British J Radiol 96(1150):20230023. https://doi.org/10.1259/bjr.20230023

Goh KH, Wang L, Yeow AYK et al (2021) Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nat Commun 12(1):711. https://doi.org/10.1038/s41467-021-20910-4

Gray M et al (2024) Measurement and mitigation of bias in artificial intelligence: a narrative literature review for regulatory science. Clin Pharmacol Ther 115:687–697. https://doi.org/10.1002/cpt.3117

Hassani BK (2021) Societal bias reinforcement through machine learning: a credit scoring perspective. AI Ethics 1:239–247. https://doi.org/10.1007/s43681-020-00026-z

Hogo MA (2020) A proposed gender-based approach for diagnosis of the coronary artery disease. SN Appl Sci. https://doi.org/10.1007/s42452-020-2858-1

Holth L, Mellstrom U (2011) Revisiting engineering, masculinity and technology studies: old structures with new openings. Int J Gend Sci Technol 3(2):313–329

Hu L, Bell D, Antani S et al (2019) An observational study of deep learning and automated evaluation of cervical images for cancer screening. JNCI: J Natl Cancer Inst 111(9):923–932. https://doi.org/10.1093/jnci/djy225

Ibrahim H, Liu X, Zariffa N et al (2021) Health data poverty: an assailable barrier to equitable digital health care. Lancet Digital Health 3(4):260–265. https://doi.org/10.1016/S2589-7500(20)30317-4

Jordansson B, Peterson H (2024) Jämställdhetsintegrering i akademin: Framgångar och fallgropar i implementeringsprocessen. Jämställdhetsmyndigheten. [Gender mainstreaming in academia: Successes and pitfalls in the implementation process. The Swedish Gender Equality Agency]. Report: 2024:2.

Joyce K, Smith-Doerr L, Alegria S et al (2021) Toward a sociology of artificial intelligence: a call for research on inequalities and structural change. Socius. https://doi.org/10.1177/2378023121999581

Khosravi P, Kazemi E, Zhan Q et al (2019) Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. NPJ Digital Med. https://doi.org/10.1038/s41746-019-0096-y

Kuhar R, Paternotte D (2017) Anti-gender campaigns in Europe. Mobilising against equality. Rowman & Littlefield International, London

Larrazabal AJ, Nieto N, Peterson V et al (2020) Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Scienced United States Am 117(23):12592–12594

Lau PL, Monomita N, Sushmita C (2023) Accelerating UN sustainable development goals with AI-driven technologies: a systematic literature review of women's healthcare". Healthcare 11(3):401. https://doi.org/10.3390/healthcare11030401

Lee MS, Guo LN, Nambdiri VE (2022) Towards gender equity in artificial intelligence and machine learning applications in dermatology. J Am Med Inform Assoc 29(2):400–403. https://doi.org/10.1093/jamia/ocab113.PMID:34151976;PMCID:PMC8757299

Liu Z (2021) Sociological perspectives on artificial intelligence: a typological reading. Sociol Compass 15:e12851. https://doi.org/10.1111/soc4.12851

Liu Y, Kohlberger T, Norouzi M et al (2019) Artificial intelligence–based breast cancer nodal metastasis detection: insights into the black box for pathologists. Arch Pathol Lab Med 143(7):859–868. https://doi.org/10.5858/arpa.2018-0147-OA

Marinucci L, Mazzuca C, Gangemi A (2023) Exposing implicit biases and stereotypes in human and artificial intelligence: state of the art and challenges with a focus on gender. AI & Soc 38:747–776. https://doi.org/10.1007/s00146-022-01474-3

Mellström U (2021) Gender studies as the political straw man. NORMA Int J Masc Stud 16(2):77–80. https://doi.org/10.1080/18902138.2021.1923899

Merriam SB (2009) Qualitative research: a guide to design and implementation. Jossey-Bass, San Francisco, CA

O'Connor S, Liu H (2023) Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities. AI Soc. https://doi.org/10.1007/s00146-023-01675-4

Oikonomou E, Williams MC, Kotanidis CP et al (2019) A novel machine learning-derived radiotranscriptomic signature of perivascular fat improves cardiac risk prediction using coronary CT angiography. Eur Heart J 40:3529–3543. https://doi.org/10.1093/eurheartj/ehz592.PMID:31504423;PMCID:PMC6855141

Olsen CM, Thompson JF, Pandeya N et al (2020) Evaluation of sex-specific incidence of melanoma. JAMA Dermatol 156(5):553–560. https://doi.org/10.1001/jamadermatol.2020.0470

European Parliament (2022) Artificial intelligence in healthcare. Applications, risks, and ethical and societal impacts. EPRS. European Parliamentary Research Service Scientific Foresight Unit (STOA).

Powell S (2018) Gender equality in academia: intentions and consequences. Int J Divers Organ Commun Nations: Annual Rev 18(1):19–35. https://doi.org/10.18848/1447-9532/CGP/v18i01/19-35

Quinn TP, Senadeera M, Jacobs S et al (2021) Trust and medical AI: the challenges we face and the expertise needed to overcome them. J Am Med Inform Assoc 28(4):890–894. https://doi.org/10.1093/jamia/ocaa268.PMID:33340404;PMCID:PMC7973477

Rajkomar A, Hardt M, Howell MD et al (2018) Ensuring fairness in machine learning to advance health equity. Annals Int Med 169(12):866–872. https://doi.org/10.7326/M18-1990

Roopaei M, Horst J, Klaas E et al. (2021) Women in ai: Barriers and solutions. In 2021 IEEE World AI IoT Congress (AIIoT) 0497–0503.

Saka E (2020) Big data and gender-biased algorithms. In: Ross K, Bachmann I, Cardo V, Moorti S, Scarcelli M (eds) The International Encyclopaedia of Gender. Media and Communication, Wiley. https://doi.org/10.1002/9781119429128

Sartori L, Theodorou A (2022) A sociotechnical perspective for the future of AI: narratives, inequalities, and human control. Ethics Inform Technol. https://doi.org/10.1007/s10676-022-09624-3

Schiebinger L, Klinge I (2015) Gendered Innovation in Health and Medicine. GENDER - Zeitschrift für Geschlecht, Kultur und Gesellschaft 7(2): 29–50. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-452079

Schiebinger L, Schraudner M (2011) Approaches to achieving gendered innovations in sience, medicine, and engineering. Interdisc Sci Rev 36(2):154–167

Sheikh H, Prins C, Schrijvers E (2023) Artificial Intelligence: Definition and Background. In Mission AI. Research for Policy, Springer, Cham

Stafanovičs A, Bergmanis T, Pinnis M (2020) Mitigating gender bias in machine translation with target gender annotations. arXiv preprint arXiv:2010.06203.

Straw EA, Stanley DA (2023) Weak evidence base for bee protective pesticide mitigation measures. J Econ Entomol 116(5):1604–1612. https://doi.org/10.1093/jee/toad118

Tomašev N, Glorot X, Rae JW et al (2019) A clinically applicable approach to continuous prediction of future acute kidney injury. Nature 572:116–119. https://doi.org/10.1038/s41586-019-1390-1

Wang T, Zhao J, Yatskar M, Chang KW, Ordonez V (2019) Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 5310–5319)

Weng SF, Reps J, Kai J (2017) Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One. https://doi.org/10.1371/journal.pone.0174944

Wesson P, Hswen Y, Valdes G et al (2022) Risks and opportunities to ensure equity in the application of big data research in public health. Annual Rev Public Health 5(43):59–78. https://doi.org/10.1146/annurev-publhealth-051920-110928

Yin RK (2018) Case study research and applications. In: Design and methods. SAGE Publications Inc

Yuan TA, Lu Y, Edwards K et al (2019) Race-, age-, and anatomic site-specific gender differences in cutaneous melanoma suggest differential mechanisms of early- and late-onset melanoma. IJERPH 16(6):908. https://doi.org/10.3390/ijerph1606090

Zajko M (2022) Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates. Sociol Compass. https://doi.org/10.1111/soc4.12962