Why does AI seem to have a problem with women?*

Silvana Badaloni[1,2,†], Antonio Rodà[1,*,†]

[1]Department of Information Engineering, via Gradenigo, 6, 35131 Padova, Italy
[2]Elena Cornaro Center on Gender Studies, University of Padova, Italy

Abstract

This position paper explores the persistent issue of gender bias in Artificial Intelligence (AI) systems. Despite increased awareness since 2018, when concerns about gender bias in AI were prominently raised, recent examples continue to demonstrate gender discrimination in AI applications. The paper examines the historical context of AI development and gender bias, analyzes current examples of gender discrimination in AI systems, and investigates the root causes of these biases. The paper presents case studies of gender bias in various AI applications, including computer vision, recommendation systems, hiring tools, and natural language processing. It also discusses the evolution of voice assistants' responses to inappropriate comments, from reinforcing stereotypes to more neutral approaches. This research contributes to the ongoing dialogue on creating more equitable and unbiased AI systems, calling for action to address gender bias in AI

Keywords

gender bias, gendered innovation, fairness, artificial intelligence, machine learning,

1. Introduction

The pervasive nature of AI in content production can fuel the generation and perpetuation of prejudices and discrimination, ultimately affecting our social and personal well-being. Among the various stereotypes perpetuated by AI, those related to gender appear to be particularly persistent and problematic. This issue was brought to the forefront in 2018 when Joe Buolamwini raised concerns about gender bias in AI systems [1]. Despite increased awareness since then, recent literature continues to provide numerous examples of gender discrimination perpetrated by AI [2].

This position paper aims to explore the complex relationship between AI and gender bias, focusing on why AI seems to have a persistent problem with women.

The relationship between stereotypes, prejudices, and discrimination forms a complex and often self-reinforcing cycle that perpetuates social inequalities. Stereotypes, defined as oversimplified beliefs about a particular group [3], serve as cognitive shortcuts but can lead to harmful generalizations. These stereotypes often form the basis for prejudices, which are negative attitudes or feelings towards individuals based on their group membership [4]. When these prejudices manifest in actions, they result in discrimination, i.e. the unfair treatment of individuals due to their group identity [5].

This cycle is self-perpetuating: existing stereotypes inform prejudices, which in turn lead to discriminatory behaviors. These discriminatory actions then reinforce the original stereotypes, creating a vicious circle. For example, gender stereotypes in STEM fields can lead to prejudiced attitudes towards women in these areas, resulting in discriminatory hiring practices. This discrimination then reinforces the stereotype that women are less suited for STEM careers [6].

Moreover, exposure to stereotypes can lead to stereotype threat, where individuals underperform in domains where their group is negatively stereotyped, inadvertently confirming the stereotype [7]. This phenomenon further entrenches existing prejudices and discrimination. Breaking this cycle requires concerted efforts at multiple levels, including education, policy changes, and individual awareness [8]. Understanding the interplay between these three phenomena is crucial for developing effective strategies to promote equality and reduce bias in various social contexts, including emerging technologies like AI.

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy
*Corresponding author.
†These authors contributed equally.
silvana.badaloni@unipd.it (S. Badaloni); antonio.roda@unipd.it (A. Rodà)

2. Gendered Machine Learning

At least since 2018, gender stereotypes and discrimination in AI have been identified in various services with an high social impact [9]: employment, medical health, mortage lending, justice systems, and in new applications of technology such as autonomous vehicles. Let's see some relevant cases in different domains.

The outcomes of three commercial gender classification systems (by Microsoft, IBM, and Face++), tested on the Pilot Parliaments Benchmark – a dataset with a balanced intersectional representation on the basis of gender and skin type – showed that all classifiers perform better on male faces than female faces (with a difference in error rate between 8.1% and 20.6%), and on lighter faces than darker faces (difference between 11.8% and 19.2%) [1]. Again with reference to computer vision techniques, gender discrimination was also found in several algorithms for pedestrian detection [10]. This task is particularly sensitive, because disparities in these algorithms could translate into disparate impact in the form of biased accident outcomes. The analysis involved the 24 top-performing methods of the Caltech Pedestrian Detection Benchmark and showed that, on average, children have higher miss rate than adults, and female a higher miss rate than males, tested on the INRIA Person Dataset [11]. However, the problem goes beyond mere computer vision, involving other applications such as automatic recommendation systems. A field test on the Facebook platform found that an advertisement promoting careers in STEM was showed more times to males than females [12]. In this case, the analysis of the outcomes showed that the bias was coded in the algorithm. Indeed, the system chose to deliver ads more to men than to women because it was designed to optimise ad delivery while keeping costs low. And the cost of an advertisement is higher if it is delivered to a woman than a man, as a consequence of the fact that women are more attractive targets as consumers (indeed, they drive 70% to 80% of all consumer purchases). Another case was reported by Amazon, that started using a hiring tool to help rank candidates using data from previous hires [13]. The system was shown to systematically downgrade female candidates and, generally, all resumes containing the word women. Interestingly, gender-based discrimination appears to be difficult to prevent due, among other causes, to a history of gender-biased hiring practices that permeate the data. This illustrates how an algorithm can potentially reinforce, and indefinitely perpetuate, already widespread discriminatory practices.

Gender bias is also an open issue for applications based on Natural Language Processing [14]. How word embedding learns stereotypes has been the focus of research on gender bias and artificial intelligence [15]. Since word embeddings are used as a knowledge base in many applications, biases in these models can propagate into many NLP applications. E.g., experiments show in many articles, papers, and websites more female names being tagged as non-person than male names, amplifying gender stereotyping [16]. In general, gender biases diffused in the text used for word-embedding – a condition often verified in textual corpora coming from writings of the last decades – are subsumed by the model: for example, words related to traditionally male professions are found closer to inherently gendered words, such as *he* or *man*, and vice versa. Techniques to reduce these biases have been recently studied [17], but the problem is not fully solved, in particular for those language that are more grammatically gendered, as Italian [18]. Reducing gender biases in textual corpora is a particularly difficult task also because automatic detection of gender bias beyond the word level requires an understanding of the semantics of written human language, which remains an open problem and successful approaches are restricted to specific domains and tasks.

3. From "I'd blush if I could" to "Don't hesitate to ask me anything that comes to your mind"

In 2019, a UNESCO report brought to light an incident involving Apple's virtual assistant, Siri, that sparked a significant debate about gender bias in AI. The report [19] highlighted a troubling response from Siri: when faced with inappropriate remarks, it would respond with the phrase, "I'd blush if I could." At first glance, this might seem like a harmless or even clever retort. However, the UNESCO report

Gender	Name	Welcome greeting
Male	Breeze	Hi, it's a pleasure to meet you. How's your day going? I can't wait to work on something beautiful with you.
Male	Arbor	Hello, I'm happy to meet you. I think we'll do a great job together. Where shall we start?
Male	Ember	Hi, I'm ready to begin. If there's anything you want me to focus on first, let me know.
Male	Cove	Hi, I just wanted to tell you that I'm excited to work with you and can't wait to get started. So, what did you have in mind?
Male	Spruce	Hi, how are you? I'm looking forward to working with you. Let's figure out where to start. What do you have in mind?
Female	Vale	Hello, it's a pleasure to meet you. If you need anything, don't hesitate to tell me. I'm here to help you.
Female	Sol	How are you? Don't hesitate to ask me anything that comes to your mind. I can start right away.
Female	Jupiter	Hi, I have a great feeling about the possibility of teaming up. How can I lend a hand?
Female	Maple	Hi, nice to meet you. I'm happy to help you achieve your goals. Let's get started!

Table 1
Welcome messages of the 9 voices (5 male and 4 female) of ChatGPT in the Italian version. The messages (originally in Italian) were transcribed on May 29, 2025.

pointed out several concerning aspects of this interaction. The response inadvertently reinforced harmful gender stereotypes by portraying Siri, with its default female voice, as a submissive and flirtatious entity. Instead of discouraging the inappropriate behavior, Siri's coy reply could be interpreted as tacitly accepting or even encouraging such comments. This not only normalized harassment in digital interactions but also missed an opportunity to promote respectful communication. Moreover, the incident raised questions about the broader implications of gendering AI assistants. By default, many of these assistants have female voices and names, potentially perpetuating the stereotype that assistants, even digital ones, should be female and accommodating regardless of how they're treated.

The Siri incident has become a pivotal example in discussions about gender bias in AI and since then, there has certainly been a greater awareness in the scientific community. Unfortunately, it's disappointing to note that gender stereotypes are still being reproduced, even by products from leading companies that, at least in their statements, place ethical aspects related to the technologies they commercialize among their priorities.

To support this affirmation, in this context we want to report a recent analysis concerning the voice interface of the latest version of ChatGPT. While at the time of the UNESCO report almost all voice assistants had pre-set female voices, in recent years the practice of giving users the option to choose the voice and gender to attribute to their voice assistant has become widespread. This choice by companies addresses some of the criticisms highlighted in the UNESCO report. Specifically, ChatGPT, at the time of writing (May 2025), offers 9 voices in the Italian version, five male and four female. Each voice, when selected, pronounces a different greeting message. Table 1 lists the greeting messages for each voice. It is evident that the messages of the male voices follow a significantly different pattern from the female voices. The male voices refer to working together, focusing on goals. The female voices, on the other hand, offer assistance, making themselves available to help. These differences clearly reproduce well-known gender stereotypes, demonstrating that while progress has undoubtedly been made, much work remains to be done

4. Causes and possible actions

Years have passed since 2018, yet many AI-based applications still display gender biases, typically disadvantaging women. This persistent issue underscores the complexity of the problem and calls for a deeper examination of its root causes and potential solutions. Generally, the main blame is placed

on the datasets used in the training phase: since these data are produced by humans within societies characterized by gender imbalances, the AI models learn these imbalances directly from data. If this is the only cause, why aren't quality datasets systematically used? Certainly, the vast amount of data needed to train large-scale models makes it uneconomical and sometimes unfeasible to curate datasets in a way that improves their quality. Often, the preference is to pre-train models with low-quality datasets, disregarding potential biases, and then limit unfair behaviors of the model during subsequent training phases. However, we believe it is then important to consider also other contributing factors.

First of all, there is a general lack of awareness among groups of designers and developers. In particular, gender issues and the potential harmful effects of stereotypes are almost ignored. It would be necessary to systematically introduce these topics in all higher education courses for engineers and programmers, in order to train a generation of technicians ready to face the challenges posed by the pervasiveness of AI tools.

Even in the presence of sufficient awareness among designers and developers, we believe there are economic reasons that discourage addressing the problem of stereotypes, particularly gender stereotypes. In practice: biased systems may be more profitable or efficient in the short term, creating a disincentive to address bias. This phenomenon can be explained by people's tendency to accommodate stereotypes already established in their society: going against stereotypes can be effortful and often can produce backlash [20]. Stereotypes can therefore be consciously used by companies to sell more or to increase the time users spend using their service. This is probably the case with the voice assistant integrated into ChatGPT, whose welcome messages were analyzed in the previous section. The designers' choice has accommodated gender expectations, with the idea that users would prefer more servile female voices and more assertive male voices. Although these choices might be economically rewarding in the short term, companies should be made aware that in the medium/long term they could instead be counterproductive because they go against the policies of many democratic governments and the sensibilities of an ever-growing portion of the population, diminishing the trust towards these technologies.

Design criteria that perpetuate stereotypes and discrimination should be made disadvantageous even in the short term, thanks to the introduction of public policies and legislative systems that discourage them. The AI Act is certainly a step forward in this direction, but it is not yet clear whether, once all its parts are operational, it will be sufficient to promote a fair and socially acceptable development of the AI-based systems.

Finally, there are various technical limitations that should be kept in mind. Current AI systems lack true understanding and often rely on statistical correlations; AI cannot differentiate between valid generalizations and unfair stereotypes. Due to feedback loops, i.e. the situations where the output of a system is fed back into it as input, the models can cause the so-called algorithmic amplification of biases. Stereotypes and fairness are human concepts, strongly culturally connoted, and are difficult to define precisely and to translate operationally in a computational manner. Even when unfair behaviors are detected, the lack of transparency and explainability of deep learning models makes it difficult to identify the causes and remedy them.

4.1. An educational approach

In the perspective of developing an Artificial Intelligence you can trust in an inclusive and ethical way, the authors have taught since A.Y. 2021-22 at the School of Engineering of the University of Padua the course "Gender Knowledge and Ethics in Artificial Intelligence" (6 CFU, 48 hour of teaching) with the aim to provide the related basic knowledge and principles in a multidisciplinary and interdisciplinary approach. The course, not compulsory for any course of study, is attended on average by about 100 students, with a gender distribution in line with that of the engineering school. As concern the contents of the course, the encounter between machines and people in contemporary society raises very central ethical questions. To this aim, it is necessary to introduce an ethical dimension applied to this discipline with a special attention to Machine Learning, facing an analysis from the point of view of gender, ethnicity, personal and social development of the ML algorithms which can lead in some cases to unfair

and discriminatory decisions. External experts have been invited to take lectures and seminars, for all the disciplinary fields outside the computer science area. The syllabus and other information on the teaching can be found on the webpage https://www.dei.unipd.it/node/35894.

In a first part of the course, particular attention has been given to some concepts concerning gender equality and gender knowledge in order to contrast stereotypes and prejudices that condition social interactions and to favor a change towards a more equitable and sustainable society. After an analysis of the differences between sex and gender [21], an attention has been given to gender statistics that characterize the world of the Academy and that have made it possible to draw up the Gender Balance. A critical reflection on the non-neutrality of knowledge and its transmission has been proposed together with an intertwining of gender, science and technology, and the centrality of a gender approach in the field of innovation to develop gendered innovation in different fields of knowledge and, in particular, in the field of Artificial Intelligence.

5. Conclusions

As noted by [22], developers of artificial intelligence are overwhelmingly male, whereas those who have reported and are seeking to address this issue are overwhelmingly female (Kate Crawford, Fei-Fei Li and Joy Buolamwini to name but a few). We strongly believe that to mitigate the presence of gender biases in computer science, diversity in the area of machine learning is essential.

It is very important to act on several perspectives. First of all, by reducing the strong under-representation of women in Artificial Intelligence. Advancing women's careers in AI, therefore, is not only a right in itself; it is essential to prevent biases and improve the efficacy of AI-based systems. Then, it is necessary to disseminate a gender culture at different levels, especially toward the younger generation, as the experience of our course has clearly shown. And an updating of training and education programmes in computer science, following a multidisciplinary approach, is perhaps one of the most promising ways to achieve these goals. Moreover, in light of the principles of gendered innovation, we believe that the study of computational techniques to analyse the presence of gender bias and mitigate its effect on outcomes represents not only an interesting problem, but first and foremost a great opportunity.

In the perspective of developing a trustworthy AI able to learn fair AI models even in spite of biased data, it is then important to address the problem of framing the landscape of gender equality and AI, trying to understand how AI can overcome gender bias and showing how an interdisciplinary analysis can help in a re-calibration of the biased tools.

Declaration on Generative AI

During the preparation of this work, the authors used Claude 3.5 in order to: text translation, grammar and spelling check. After using these tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: Conf. on fairness, accountability and transparency, 2018, pp. 77–91.

[2] H. Kotek, R. Dockum, D. Sun, Gender bias and stereotypes in large language models, in: Proceedings of the ACM collective intelligence conference, 2023, pp. 12–24.

[3] G. W. Allport, K. Clark, T. Pettigrew, The nature of prejudice, Addison-wesley publishing company Cambridge, MA, 1954.

[4] J. F. Dovidio, M. Hewstone, P. Glick, V. M. Esses, Prejudice, stereotyping and discrimination: Theoretical and empirical overview, Prejudice, stereotyping and discrimination 12 (2010) 3–28.

[5] D. Pager, H. Shepherd, The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets, Annu. Rev. Sociol 34 (2008) 181–209.
[6] S. Cheryan, S. A. Ziegler, A. K. Montoya, L. Jiang, Why are some stem fields more gender balanced than others?, Psychological bulletin 143 (2017) 1.
[7] C. M. Steele, J. Aronson, Stereotype threat and the intellectual test performance of african americans., Journal of personality and social psychology 69 (1995) 797.
[8] E. L. Paluck, D. P. Green, Prejudice reduction: What works? a review and assessment of research and practice, Annual review of psychology 60 (2009) 339–367.
[9] A. Nadeem, B. Abedin, O. Marjanovic, Gender Bias in AI: A Review of Contributing Factors and Mitigating Strategies, in: Proc. of the 31st Australian Conference on Information Systems, New Zealand, 2020, p. 12.
[10] M. Brandao, Age and gender bias in pedestrian detection algorithms, in: Workshop on Fairness Accountability Transparency and Ethics in CV at CVPR 2019, 2019.
[11] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), volume 1, Ieee, 2005, pp. 886–893.
[12] A. Lambrecht, C. Tucker, Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads, Management science 65 (2019) 2966–2981.
[13] J. Dastin, Amazon scraps secret ai recruiting tool that showed bias against women, in: Ethics of Data and Analytics, Auerbach Publications, 2018, pp. 296–299.
[14] J. Doughman, W. Khreich, M. El Gharib, M. Wiss, Z. Berjawi, Gender bias in text: Origin, taxonomy, and implications, in: Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing, 2021, pp. 34–44.
[15] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, Advances in neural information processing systems 29 (2016).
[16] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, W. Y. Wang, Mitigating gender bias in natural language processing: Literature review, in: Proc. of the 57th Annual Meeting of the Ass. for Comput. Linguistics, 2019, pp. 1630–1640.
[17] S. L. Blodgett, S. Barocas, H. Daumé III, H. Wallach, Language (technology) is power: A critical survey of "bias" in nlp, arXiv preprint arXiv:2005.14050 (2020).
[18] D. Biasion, A. Fabris, G. Silvello, G. A. Susto, Gender bias in italian word embeddings., in: CLiC-it, 2020.
[19] M. West, R. Kraut, H. Ei Chew, I'd blush if I could: closing gender divides in digital skills through education, Unesco, 2019.
[20] L. A. Rudman, J. E. Phelan, Backlash effects for disconfirming gender stereotypes in organizations, Research in organizational behavior 28 (2008) 61–79.
[21] A. Viola, Il sesso è (quasi) tutto: Evoluzione, diversità e medicina di genere, Feltrinelli Editore, 2022.
[22] S. Leavy, U. C. Dublin, Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning, in: Proc. of the ACM/IEEE 1st International Workshop on Gender Equality in Software Engineering, Gothenburg, Sweden, 2018.