ANALYSIS

**Monique Munarini**
January 2025

# Multi-stakeholder guidelines on how to address gender bias in AI systems

Friedrich Ebert Stiftung

Competence Centre on the Future of Work

## Imprint

**Published by**
Friedrich-Ebert-Stiftung e.V.
Godesberger Allee 149
53175 Bonn, Germany
info@fes.de

**Issuing Department**
Competence Centre on the Future of Work
Cours Saint Michel 30a, 1040 Brussels, Belgium

For more information about the Competence Centre
on the Future of Work, please consult:
https://www.futureofwork.fes.de

**Responsibility for Content and Editing**
Dr. Inga Sabanova
inga.sabanova@fes.de

**Design/Layout**
pertext | corporate publishing
www.pertext.de



January 2025


ISBN 978-3-98628-689-7

Further publications of the Friedrich-Ebert-Stiftung can be found here:
↗ www.fes.de/publikationen

Friedrich Ebert Stiftung

**Monique Munarini**
January 2025

# Multi-stakeholder guidelines on how to address gender bias in AI systems

# Content

# Summary

As artificial intelligence (AI) systems increasingly influence critical sectors such as healthcare, employment, education and law enforcement, concerns around bias – especially gender bias – have come to the forefront. Gender bias in AI not only reflects but can escalate existing inequalities, raising significant ethical, legal and societal issues. This policy paper examines the impact of gender bias in AI systems and presents comprehensive guidelines for addressing it through a socio-technical lens. By focusing on different stages of the AI lifecycle, the paper provides actionable recommendations for various stakeholders, including developers, deployers, users and regulators. Therefore, the aims of this document are to:

→ raise awareness about the escalation of gender bias when using AI systems in the decision-making process;

→ outline the challenges and opportunities of incorporating a socio-technical approach to tackle gender bias issues in AI systems;

→ provide a set of recommendations to key stakeholders from a socio-technical perspective on how to identify and prevent the reproduction of gender bias.

The first part of this paper introduces the challenge of gender bias in AI systems, highlighting its root causes, including biased training data, lack of diverse representation in AI development teams, and insufficient consideration of the social impacts of AI. Gender bias can have far-reaching consequences, from discriminatory hiring practices to skewed medical diagnoses. This cannot be tackled from a technical standpoint alone. The paper advocates a socio-technical approach, which views AI systems as both technical and social constructs. This approach emphasises the need for interdisciplinary collaboration and the inclusion of diverse voices, particularly those of marginalised groups, throughout the AI lifecycle.

The second part of this paper provides tailored guidance for different stakeholders. For developers, it outlines best practices to avoid gender bias in data and algorithms. For deployers, including small and medium-sized enterprises (SMEs), it offers strategies to ensure that AI systems are used equitably to uphold fundamental rights, avoiding potential harms from gender-biased outcomes. Users are provided with instructions on how to interact with AI systems in a way that recognises potential biases and to take actions to flag potential issues. Finally, for regulators, the paper suggests policy recommendations, drawing on frameworks such as the EU AI Act and the UNESCO AI Principles.

This policy paper also includes cases of how bias can impact the decision-making process and the importance of implementing the guidelines set out in this document.

# Introduction

Artificial intelligence (AI) is now accessible to small and medium-sized enterprises through "AI as a service" offerings (Cobbe and Singh, 2021), reshaping industries from healthcare to finance. The democratisation of access to AI has enabled small and medium-sized enterprises to integrate AI-driven technologies into their operations, in search of a competitive advantage. McKinsey's 2024 global report highlights that AI adoption has surged by 50% in the past six years, confirming its role as a transformative force in society (McKinsey & Co, 2024).

There are great advantages to incorporating AI systems in supporting decision-making processes to augment human capabilities where AI systems can provide predictions or detect correlations that might pass unnoticed to humans (Cummings, 2004). However, alongside the excitement around AI hype (Fishburne, 2024), there is growing concern about its ethical, social and legal implications, particularly regarding gender bias. Instead of simply trying to navigate the AI hype in an ethical and lawful manner, the first question that needs to be asked is not **if AI systems can be incorporated into a process or product, but if AI should be used in the first place**. As Gabriella Ramos (2024) noted at UNESCO's Women 4 Ethical AI conference, "if we can make it for women, we can make it for all". Gender bias does not only relate to the binary definition of men and women, but accounts for the perspective of using men as a default for system design (Perez, 2019), excluding other genders, including the female gender, which makes up more than half of the world's population. This bias is not just an ethical issue but a systemic one, perpetuated throughout the AI lifecycle.

The reconciliation of ethical and legal regulations such as the UNESCO Recommendation on the Ethics of Artificial Intelligence (2021)[1] and the European Union regulation on artificial intelligence, or EU AI Act (AIA) (European Parliament, 2024) emphasise the need[2] for AI literacy, urging actors throughout the lifecycle to minimise potential discriminatory outcomes. A socio-technical approach – accounting for human, organisational and technical factors – is essential to address these challenges and foster responsible development and deployment of AI systems.

Box 1

## *Auditing AI models*

*Researchers from Stanford Law School (Haim et al., 2024) conducted an audit of OpenAI and Google AI models to examine the impact of names on the outcomes generated by language models. The study involved requesting advice on key policy-relevant issues, including socio-economic status, intellectual capabilities, electability and employability. The researchers found that names perceived as female consistently produced worse outcomes than those perceived as male.*

These guidelines aim to promote AI literacy, by identifying these AI actors and where they should intervene, and recommending reasonable efforts to avoid the escalation of gender bias in AI systems.

1 According to UNESCO (2022), "AI actors should make all reasonable efforts to minimise and avoid reinforcing or perpetuating discriminatory or biased applications and outcomes throughout the life cycle of the AI system to ensure fairness of such systems" (*UNESCO Recommendation on the Ethics of Artificial Intelligence*, p. 20).

2 Recital 20 of the EU AI Act reasons that AI literacy should equip providers, deployers and affected persons with the necessary notions to make informed decisions regarding AI systems.

# Overview of gender bias in AI systems

## Understanding the AI lifecycle and mapping stakeholders

According to the OECD (2024), the AI lifecycle comprises key phases: **design, data collection and processing, model building and validation, deployment, operation and monitoring, and retirement and decommission. The lifecycle begins with the plan & design phase**, where developers and designers conceptualise the purpose, scope and ethical guardrails for the AI system. Here, product managers, ethicists and user experience (UX) designers have the opportunity to embed ethical principles, often known as ethics-by-design, ensuring that the AI aligns with upholding ethical values as core objectives from the outset.

During **data collection and processing**, dataset creators/designers and curators select, clean and prepare data. These stakeholders must ensure datasets are diverse, representative of the end-user population, and scrutinised. In the **model building and validation** phase, machine-learning engineers and data scientists train and test models, making critical choices about algorithms, parameter settings and evaluation metrics. These choices should include performance assessments across diverse demographics. The sequence of design, data and models is context-dependent and needs a thorough study about the possible impacts of this system once placed in the real world (post-deployment).

As the system moves to **deployment**, compliance officers, companies who incorporate AI systems in their business operations and policymakers play a significant role in enforcing regulatory frameworks, ensuring the model is fit for use and adheres to regulatory standards. Post-deployment, in the **operation and monitoring phase**, public auditors, user advocacy groups and end-users provide essential feedback on AI performance. This stage also involves policymakers, who may adjust regulations based on observed issues and societal impacts. The ultimate phase in the AI lifecycle is the **decommissioning** of this system. The termination of an AI system, for example, can follow any of the other phases, whenever its operation loses its purpose or fails the monitoring process.

These phases encapsulate the journey from initial AI conception to active use and long-term oversight. **The ethical impact of AI systems should be assessed in each phase of the AI lifecycle**, as all these phases take place iteratively and do not necessarily follow this exact sequence.

## Gender bias in AI systems

There are two distinct but interrelated definitions of bias that are crucial for understanding gender bias in AI systems: **human bias** and **technical bias**.

The first, **human bias** (also known as cognitive bias), refers to systematic deviations in human judgement from probabilistic expectations (Tversky et al., 1982). Rooted in cultural, historical and societal contexts, cognitive bias can create prejudices against specific groups, including women (Cirillo & Rementeria, 2022). In the case of gender bias, these factors create unwanted preferences against one group, considering their unique attributes, resulting in the basis of discrimination (Munarini, 2022). This is even more critical in the political, educational and economic fields. Furthermore, even when considering their physical integrity, women are disadvantaged due to biases. For instance, the 2023 Gender Social Norms Index (GSNI) compiled by the United Nations Development Programme (UNDP, 2023) revealed that **nearly 9 out of 10 men and women globally are biased against women**. AI systems were initially adopted in decision-making to counteract human bias, under the assumption that machines are inherently impartial (Raso et al., 2018, 7).

The second type, technical bias (or statistical bias), arises from deviations in expected outcomes (Ferrer et al, 2021). In a simple example, in a pool full of blue balls, finding a red ball would be the deviation, but it could help classify the components of the pool and understand that there can be

**Box 2**

### *Gender bias in AI systems in practice*

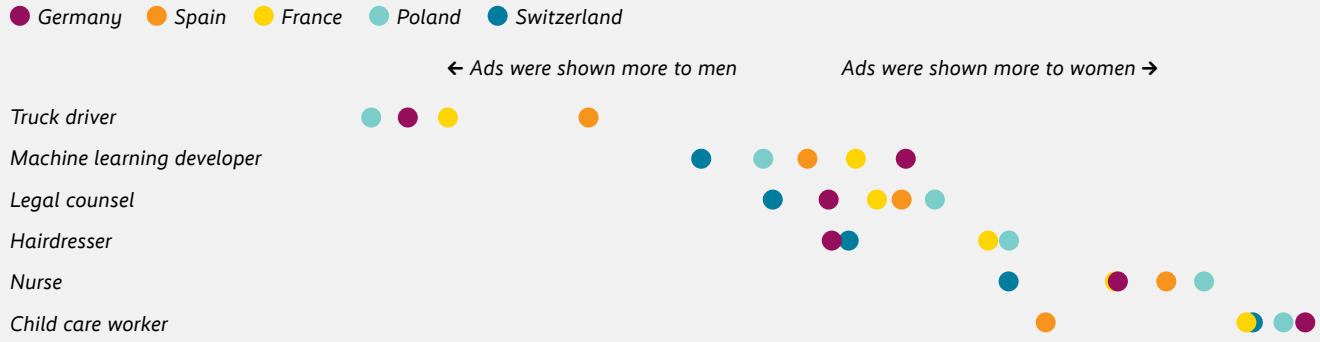
*In 2014, Amazon's AI-based CV screening system was found to be systematically deleting applications from women. The system had been trained on a decade's worth of employee data, which was predominantly male, reflecting historical biases. Consequently, the AI system "learned" that female candidates were unsuitable for roles at Amazon (Dastin, 2018).*

## *Optimisation tool for job delivery*

**Fig. 1**

*A report by AlgorithmWatch (2019) reproduced in Europe a study done in the US on job ad delivery tools. The findings revealed significant gender bias in job ad delivery algorithms used by Meta (formerly Facebook). These tools, designed to optimise engagement, perpetuated stereotypes by showing truck driver ads predominantly to men and care-related job postings to women, even without explicit gender targeting by advertisers. In Germany, for instance, truck driver ads were shown to only 386 female users, compared with 4,864 male users, demonstrating the algorithm's reinforcement of biased assumptions.*

*When deciding who to show an ad to, Facebook relies on gross stereotypes*
*We bought ads for six different job offers in five countries. This is how Facebook optimized the ad impressions, bases on gender*




Based on 102,472 ad impressions between 27 August and 23 September.
Source: AlgorithmwatchWatch.

more than just blue balls in that pool or that most of the balls in that pool can be categorised as blue. In machine learning, bias is inferred from statistical bias linked to errors that consistently misclassify data, often resulting from flaws in training, modelling or system usage (Crawford, 2021, 134). Bias is often unintentional, but technical and human biases frequently overlap, particularly when designers fail to recognise or mitigate their own biases (Coeckelbergh, 2020, 125). As AI systems increasingly influence decision-making processes, their unchecked biases can amplify discriminatory outcomes at scale. Kate Crawford aptly summarised this challenge, noting that "in the rush to arrive at a narrow technical solution of statistical bias as though that is a sufficient remedy for deeper structural problems" (Crawford, 2021, 135). While technical solutions can address statistical inaccuracies, they are insufficient without addressing the human bias embedded in AI systems.

A socio-technical approach to tackling issues such as gender bias in AI systems calls for decisions taken at each step of the AI lifecycle by key actors to consider not only technical features, but the social components of AI systems, i.e. the people who are creating the system, deploying it or being affected by it.

## Socio-technical approach to mitigate gender bias in AI systems: an introduction

As demonstrated in the previous section, gender bias in AI systems can emerge from both social and technical factors. The societal impact of AI systems extends far beyond data, algorithms or data centres. **Decisions made during design and development – such as whom the systems are created for and whose experiences are prioritised – highlight power asymmetries that significantly influence societal outcomes** (Perez, 2019, 176). Many unintended consequences of AI systems can stem from seemingly minor initial decisions that, while fitting the system's design parameters, can escalate bias over time (Pedreschi et al., 2023). Given the inherently socio-technical nature of AI systems, addressing challenges such as gender bias requires an approach that integrates these dimensions. A socio-technical perspective offers a

**Box 3**

### *The socio-technical approach in practice*

*UNESCO's Global AI Ethics and Governance Observatory (2024a) exemplifies a socio-technical approach by sharing country profiles on AI governance efforts based on publicly available data and its Readiness Assessment Methodology (RAM). This tool goes beyond AI's design, development and deployment to consider its broader socio-technical context. Factors such as access to electricity (Brazil), gender gaps in STEM careers (Morocco) and the influence of data protection regulations on AI governance (Gabon) are integral to these profiles, reflecting how AI systems both shape and are shaped by their institutional, geographical and cultural environments.*

pathway to developing responsible, ethically aligned AI systems by examining how systems are designed, who is involved in their creation, and why specific design choices are made (Dignum, 2019, 48). While technical experts focus on ensuring the system functions accurately, they should not independently decide on trade-offs that affect accuracy and other critical outcomes. Often the process is fragmented: systems are developed first, and only later are experts from social disciplines invited to address the ethical or legal implications. This responsibility is not limited to individual designers or confined to a single stage of the AI lifecycle. Instead, it demands an ongoing, collaborative effort across disciplines working towards a shared objective (OECD, 2024; Dignum, 2023).

As the adoption of AI systems accelerates, fuelled by the ongoing AI hype, it becomes imperative to move beyond surface-level fixes and implement actionable guidelines to address gender bias. The next section, **Navigating the AI hype: guidelines to address gender bias in AI systems**, provides a practical roadmap for stakeholders to confront these challenges effectively and build reliable AI systems in a responsible way.

# Navigating the AI hype: guidelines to address gender bias in AI systems

These guidelines are an effort to develop a holistic approach with a socio-technical perspective promoting the responsible development and deployment of AI systems. While they can be addressed individually, they would be more effectively incorporated as part of a comprehensive strategy. They serve as a starting point for further discussion, with the ultimate goal of highlighting issues that might otherwise be overlooked in the AI lifecycle. One of the key objectives of this document is to encourage further research to create more comprehensive actions for mitigating gender bias in AI systems.

The initial section on the AI lifecycle has helped map key stakeholders who could benefit from guidance from a socio-technical perspective to address gender bias in AI systems. The terminology used aligns with the EU AI Act (AIA), which came into force on 1 August 2024 (European Parliament, 2024). According to the EU AI Act, providers are those involved in the development or commercialisation of an AI system or a general-purpose AI model (Article 3(3) AIA). Deployers are public or private entities that use AI systems professionally (Article 3(4) AIA). Users are individuals who may be affected by the outcomes of AI systems once deployers use them, or those who use AI systems for personal purposes, such as students using ChatGPT or other generative AI to assist with homework, or travellers using translation tools to access information in a foreign country.

## Guidance for providers

**Staff diversity is NOT just adding more women:** providers should set clear goals and metrics to measure the diversity of technology teams working on AI, taking into account the objectives of the AI system or model under development. Establishing targets and incentives for **workforce diversity** – going beyond simply counting the number of people hired and focusing on those actively engaged in the design and development process – is essential to avoid "pink-washing" or "ethics-washing". **Leadership diversity** is crucial to ensure that AI development isn't driven by a homogenous group that may overlook the needs of non-hegemonic voices, leading to "unreflective data decisions" (Simson et al., 2024).

Contextual knowledge plays a crucial role in mitigating gender bias throughout the AI lifecycle. For this reason, engaging **not only technical experts** but social scientists, ethicists and community representatives ensures the perspectives of those who will be directly affected by the system are embedded in its design and development from the outset.

Furthermore, deployers acquiring AI systems should have access to **demographic information** about the teams involved in building the systems to assess potential sources of bias embedded within them.

**All data matters:** providers should be encouraged to create public repositories of datasets that are not just gender-sensitive, but also acknowledge the intersectionality of gender with other factors such as race, socio-economic status, disability, etc. Providing an **open forum for feedback on datasets** could further help mitigate unintended bias in the data. These public repositories are an important accountability and transparency measure, as they can be open to feedback from third-party assessments to ensure that the data reflect real-world diversity and lived experiences. An internal strategy would be the creation of a **monitoring working group** to assess the overall diversity and intersectional perspectives of each phase of the AI lifecycle, from design to placing on the market.

**Bias mitigation techniques:** several machine-learning techniques are gaining attention as part of bias mitigation strategies in AI development. While the selection and deployment of bias metrics remains controversial, some techniques are becoming increasingly relevant within the machine-learning community to ensure not only the mitigation of bias, but also the enhancement of privacy protection. The United Nations University (UNU) recently published a policy brief recommending the **use of synthetic data** to train AI models aligned with these purposes (Wilde et al., 2024).

## Guidance for deployers

**Implementation of contestability measures:** a key requirement for integrating AI systems into decision-making processes is the establishment of clear contestability mechanisms. End-users, or those directly impacted by AI outcomes, **must have access to explanations about the role of AI in decision-making and information that enables them to challenge decisions**. Such

measures promote greater transparency and accountability while meeting legal obligations.

**Ongoing performance monitoring:**[3] one-off validation is insufficient to prevent discriminatory outcomes or disproportionate impacts on vulnerable groups. A strategic plan for continuous performance evaluation, incorporating internal or external audits, ensures equitable accuracy across different demographic groups. For SMEs and startups, cost-effective alternatives such as public red-teaming – where civil society organisations and research centres test AI systems to uncover vulnerabilities – can provide robust accountability without an excessive financial burden. Providers such as OpenAI, IBM and Anthropic have adopted this practice internally. A one-time validation effort is not enough to ensure that the system does not disproportionately affect vulnerable groups, leading to discriminatory outcomes. The design of a strategic plan to assess the system's performance, taking into account internal or external auditing mechanisms, supports a distributed accuracy across different target demographics in the system's deployment. In the specific context of SMEs and startups, a possible alternative to implementing this initiative without excessive implementation costs and human oversight of the AI system in use is the adoption of public red-teaming. Red-teaming in AI is a recognised practice that simulates the deployment of an AI system in the real world to uncover vulnerabilities and flaws. This practice is already implemented by some providers of AI systems such as OpenAI (2024), IBM (2024) and Anthropic (2024) as an internal accountability measure. Public red-teaming (Humane, 2024) usually involves civil society organisations or research centres that muster experts and non-experts to test AI systems in order to assess ethical and privacy flaws.

**Creation of an AI committe**e: public or private organisations can incorporate AI boards or committees in their governance pipeline. These are responsible for accountability strategies in the development and deployment of the AI system regarding bias mitigation and other ethical concerns. The experts appointed to those boards or committees would be responsible for impact and fundamental rights assessments. The creation of a committee to oversee AI systems would also mean managing an **inventory** with the strategies implemented in the AI lifecycle.

**AI "driving licence":** under the EU AI Act, AI literacy is a shared responsibility of providers and deployers. Personnel should receive training akin to obtaining a driving licence, raising awareness of common biases and associated risks and ensuring responsible use of AI systems. Regularly renewing this "licence" would **test for overreliance on AI and provide an opportunity to gather feedback on system performance**.

## Guidance for policymakers

Policymakers have a unique position in the AI lifecycle. As stakeholders, they act as a **convergence point, responsible for fostering the participation of civil society organisations (CSOs) and representatives of vulnerable groups**, providing the resources they need to make a meaningful contribution to exploring how AI governance can be more inclusive and less biased.

**Development of working groups for shared dialogue:** this should take place between all stakeholders involved in the AI lifecycle, with a constant monitoring process to improve the accountability of the AI system in use and future projects involving other AI-powered solutions. Gender mainstreaming strategies are key in designing these projects. These could be in the form of a **forum for community feedback** on the work done by AI boards (as suggested above), marshalling different expertise with the common goal of developing and deploying reliable AI systems.

**Designing strategies for inclusive participation of external stakeholders:** to foster AI literacy among citizens subject to AI outcomes, initiatives must address the digital divide, especially among vulnerable groups (UNESCO, 2024). **Programmes aimed at removing barriers** for migrants, older adults, people with disabilities, survivors of gender-based violence and even children are essential. These **efforts empower individuals to engage in discussions about AI governance and actively participate in shaping its strategies**. It is not possible to talk about AI, gender bias or discrimination if regular people are left behind in the digital revolution. Apart from enabling people to access services and use products, one of the aims of mitigating bias is to avoid discriminatory outcomes.

**Clear governance mechanisms:** as the EU AI Act unfolds, policymakers should establish governance frameworks with tangible key performance indicators (KPIs) tailored to regional and local contexts. These benchmarks need not be limited to metrics, but could include other measures to guide providers and deployers in developing oversight strategies. Resources such as the US Department of State's *Risk Management Profile: Artificial Intelligence and Human Rights* (2024) offers practical guidelines for aligning AI systems with international human rights standards.

**Whistleblower protection:** policymakers should create policies ensuring safe reporting of concerns without risk of retaliation or restrictive agreements from individuals collaborating with providers or deployers, not only regarding gender bias in AI systems, but in support of ethical and safe AI governance. As highlighted in UNE-

3 Aligned with recital 65 AIA.

SCO's *Recommendation on the Ethics of Artificial Intelligence* (2021) (recommendation 43), such measures are critical to improving accountability in the AI lifecycle.

## Guidance for users/affected persons

**Understanding your rights:** the mainstreaming of AI systems presents opportunities to learn about the risks and benefits of AI. Engaging with local initiatives or online courses can help individuals understand how AI impacts daily life and raise awareness about misuse. This knowledge encourages **informed participation in AI governance** and boosts regulation discussions.

**Seeking support:** cross-cultural dynamics increase the complexity of understanding and addressing gender bias in AI systems. At the regional and international level, there are many civil society associations that are involved in **providing support to victims** of different types of discrimination, including algorithmic discrimination arising from gender bias. Seeking support can enhance understanding and facilitate community discussions about the societal implications of AI.

**Engaging in the feedback loop:** individuals do not need to be experts to contribute to discussions about AI systems. Users should feel empowered to report concerns about AI outcomes through available feedback mechanisms. If feedback channels are inaccessible, this issue itself is valuable information to highlight. When something seems wrong, **the most important thing is not to ignore it**. Public forums, red-teaming exercises, focus groups and workshops are also excellent opportunities for users to voice their concerns and contribute to AI improvement efforts. There are many sources online or at local hubs such as community centres or collectives engaged in such projects.

# Conclusion

These guidelines were developed to enlighten the experience of different stakeholders in the AI lifecycle when trying to understand their roles in the mitigation of gender bias in AI systems. Recognising that much more needs to be done to effectively mitigate gender bias in AI systems is crucial. Simply selecting the right metric or individualistic solution is insufficient when proposing holistic strategies that encompass the AI lifecycle. Adopting a socio-technical approach emphasises that addressing bias requires more than technical solutions: it demands connected strategies involving all actors in the AI lifecycle. The AI literacy initiatives outlined here extend beyond compliance with the EU AI Act; they promote a responsible AI future aligned with shared ethical frameworks.

# Appendix

The EU AI Act (AIA) (European Parliament, 2024) adopted the updated definition from the OECD AI principles (2024): an AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.

AI Hype: phenomenon of incorporating AI as a service or at least advertising that AI is part of a product, such as AI-powered solutions present in many products from online translators to toothbrushes.

AI lifecycle: AI system lifecycle phases involve (i) "design, data and models", which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; (ii) "verification and validation"; (iii) "deployment"; and (iv) "operation and monitoring". These phases often take place iteratively and are not necessarily sequential. The decision to retire an AI system from operation may occur at any point during the operation and monitoring phase.

AI literacy: AI literacy should equip providers, deployers and affected persons with the necessary notions to make informed decisions regarding AI systems. Those notions may vary with regard to the relevant context and can include understanding the correct application of technical elements during the AI system's development phase, the measures to be applied during its use, the suitable ways in which to interpret the AI system's output, and, in the case of affected persons, the knowledge necessary to understand how decisions taken with the assistance of AI will have an impact on them (recital 20 AIA).

Algorithmic bias: algorithmic bias happens when an algorithm, or a set of computer instructions, unfairly discriminates against certain people or groups (UNESCO, 2024b).

Gender bias: prejudiced actions or thoughts based on the gender-based perception that women are not equal to men in rights and dignity (EIGE, 2026).

Gender mainstreaming: gender mainstreaming involves applying a gender equality perspective in each phase of the policy-making cycle as well as all areas within policies and processes such as procurement or budgeting (EIGE, 2024).

Gender Social Norms Index (GSNI): the UNDP Index (2023) quantifies biases against women, capturing people's attitudes on women's roles along four key dimensions: political, educational, economic and physical integrity. The 2023 GSNI, covering 85 percent of the global population, reveals that close to 9 out of 10 men and women are fundamentally biased against women.

# References


AlgorithmWatch. (2019). *Automated discrimination on Facebook and Google: Systems influence job ads shown to women and men differently.* Retrieved 20 November 2024, from https://algorithmwatch.org/en/automated-discrimination-facebook-google/

Anthropic. (2024). *Challenges in red-teaming AI systems.* Retrieved 21 November 2024, from https://www.anthropic.com/news/challenges-in-red-teaming-ai-systems

Cirillo, D., & Rementeria, M. J. (2022). Bias and fairness in machine learning and artificial intelligence. *Sex and gender bias in technology and artificial intelligence* (pp. 57–75). Academic Press.

Cobbe, J., & Singh, J. (2021). Artificial intelligence as a service: Legal responsibilities, liabilities, and policy challenges. *Computer Law & Security Review*, 42, 105573.

Coeckelbergh, M. (2020). *AI ethics.* MIT Press.

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

Cummings M. (2004). Automation bias in intelligent time critical decision support systems. *AIAA 1st Intelligent Systems Technical Conference.* Reston, VA: American Institute of Aeronautics and Astronautics.

Dignum, V. (2023, January). Responsible Artificial Intelligence – From Principles to Practice: A Keynote at TheWebConf 2022. *ACM SIGIR Forum* (Vol. 56, No. 1, pp. 1–6). New York, NY, USA: ACM.

— (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way* (Vol. 2156). Cham: Springer.

Dastin, J. (2018), Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women, Reuters. Retrieved 15 November 2024, from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

European Institute for Gender Equality (EIGE) (2016). Gender bias. European Institute for Gender Equality. Retrieved 21 November 2024, from https://eige.europa.eu/publications-resources/thesaurus/terms/1320?language_content_entity=en

— (2024). *Gender mainstreaming*. European Institute for Gender Equality. Retrieved 21 November 2024, from https://eige.europa.eu/gender-mainstreaming?language_content_entity=en

European Parliament (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). (2024), 1–144.

Fishburne, T. (2024, March 24). *The potential of AI [Cartoon].* Marketoonist. https://marketoonist.com/2024/03/potential-of-ai.html

Ferrer, X., Van Nuenen, T., Such, J. M., Coté, M., & Criado, N. (2021). Bias and discrimination in AI: a cross-disciplinary perspective. *IEEE Technology and Society Magazine*, 40(2), 72-80.

Haim, A., Salinas, A., & Nyarko, J. (2024). What's in a Name? Auditing Large Language Models for Race and Gender Bias. *arXiv preprint arXiv:2402.14875.*

Humane Intelligence (2024). *Generative AI Red Teaming Challenge*. Retrieved 21 November 2024, from https://www.humane-intelligence.org/grt

IBM (2024). *What is Red Teaming?* IBM. Retrieved 21 November 2024, from https://www.ibm.com/think/topics/red-teaming

McKinsey & Company (2024). *The state of AI 2024: Trends and developments in artificial intelligence.* Retrieved from https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

Munarini, M. (2022). New perspectives on the mitigation of gender bias in AI by EU regulations. *Peace Human Rights Governance,* 6 (Peace Human Rights Governance 6/2), 111–136.

OpenAI (2024). *Red teaming network.* Retrieved 21 November 2024, from https://openai.com/form/red-teaming-network/

Organisation for Economic Co-operation and Development (OECD) (2024). OECD AI principles: Recommendations of the Council on Artificial Intelligence. Retrieved 20 November 2024, from https://oecd.ai/en/ai-principles

Pedreschi, D., Pappalardo, L., Baeza-Yates, R., Barabasi, A. L., Dignum, F., Dignum, V., ... & Vespignani, A. (2023). Social AI and the challenges of the human-AI ecosystem. *arXiv preprint arXiv:2306.13723.*

Perez, C. C. (2019). *Invisible Women: Exposing Data Bias in a World Designed for Men.* 1st edition. Vintage.

Ramos, G. (30 October 2024). Quote from the Women 4 Ethical Artificial Intelligence Conference. *Women 4 Ethical Artificial Intelligence Conference.* UNESCO. Retrieved from https://www.unesco.org/en/articles/women-4-ethical-artificial-intelligence-conference

Raso, F. et al. (2018). Artificial Intelligence & Human Rights: Opportunities & Risks. *Berkman Klein Center for Internet & Society Research Publication*, Harvard University's DASH repository. Retrieved 20 November 2024, from: http://nrs.harvard.edu/urn-3:HUL.InstRepos:38021439

Simson, J., Fabris, A., & Kern, C. (2024, June). Lazy data practices harm fairness research. *2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 642–659).

Tversky, A., Kahneman, D., & Slovic, P. (1982). Judgment under uncertainty: Heuristics and biases (pp. 3–20).

United Nations Development Programme (UNDP) (2023). *Gender Social Norms Index (GSNI) 2023.* UNDP. Retrieved 6 November2024, from https://hdr.undp.org/content/2023-gender-social-norms-index-gsni#/indicies/GSNI

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2024a). *Ethics of artificial intelligence: Global hub.* UNESCO. Retrieved 5 November 2024, from https://www.unesco.org/ethics-ai/en/global-hub

— (2024b). *Challenging systematic prejudices: An investigation into bias against women and girls in large language models* (CI/DIT/2024/GP/01). Retrieved 5 November 2024, from https://unesdoc.unesco.org/ark:/48223/pf0000388971

— (2021). *Recommendation on the ethics of artificial intelligence.* Retrieved 5 November 2024, from https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

U.S. Department of State (2024). *Risk management profile for AI and human rights.* U.S. Department of State. Retrieved 21 November 2024, from https://www.state.gov/risk-management-profile-for-ai-and-human-rights/

Wilde, P., Arora, P., Buarque, F., Chin, Y. C., Thinyane, M., Stinckwich, S., Fournier-Tombs, E., & Marwala, T. (2024). *Recommendations on the use of synthetic data to train AI models.* United Nations University (UNU). https://unu.edu/publication/recommendations-use-synthetic-data-train-ai-models

## About the Author

**Monique Munarini** is a trained lawyer with a double master's degree in international relations from the University of Padova (IT) and Law, Economics and Management from the University of Grenoble-Alpes (FR). She is currently a PhD candidate in the Italian National PhD in AI at the University of Pisa (IT) where her research focuses on developing equitable AI audits from a feminist perspective. At the same time, she also works as policy analyst in AI Governance for International Organisations and Global Civil Society Organisations.

# Multi-stakeholder guidelines on how to address gender bias in AI systems

As artificial intelligence (AI) systems increasingly influence critical sectors such as healthcare, employment, education and law enforcement, concerns around bias – especially gender bias – have come to the forefront. Gender bias in AI not only reflects but can escalate existing inequalities, raising significant ethical, legal and societal issues. This policy paper examines the impact of gender bias in AI systems and presents comprehensive guidelines for addressing it through a socio-technical lens. By focusing on different stages of the AI lifecycle, the paper provides actionable recommendations for various stakeholders, including developers, deployers, users and regulators. Therefore, the aims of this document are to:

→ raise awareness about the escalation of gender bias when using AI systems in the decision-making process; outline the challenges and opportunities of incorporating a socio-technical approach to tackle gender bias issues in AI systems;

→ provide a set of recommendations to key stakeholders from a socio-technical perspective on how to identify and prevent the reproduction of gender bias.

Friedrich Ebert Stiftung