Bialystok Legal Studies
Białostockie Studia Prawnicze
2026 vol. 31 no. 1

sciendo

DOI: 10.15290/bsp.2026.31.01.08

Received: 07.08.2025
Accepted: 13.12.2025

Zeynep Ayata
Istanbul Policy Center, Sabancı University, Turkey
zeynep.ayata@sabanciuniv.edu

Gender Bias in AI Systems: A Critical Analysis of Regulatory Frameworks and Policy Responses

Abstract: The rapid proliferation of artificial intelligence systems has exposed pervasive gender biases that reflect and amplify existing societal inequalities, posing significant threats to gender equality and women's fundamental rights. This article examines gender bias in AI systems through both theoretical and regulatory lenses, analysing how these biases manifest and can be addressed through comprehensive policy frameworks. The first section provides a systematic literature review exploring how bias becomes embedded in algorithmic systems through biased training data, algorithmic design choices, and broader cultural contexts. The second section examines policy responses, comparing UNESCO's comprehensive recommendations with the European Union's Artificial Intelligence Act and referencing the Council of Europe Framework Convention on Artificial Intelligence. This analysis reveals a significant disconnect between aspirational frameworks and practical implementation, demonstrating that existing regulatory approaches inadequately address gender bias in AI and highlighting the urgent need for comprehensive integration of gender equality considerations into AI governance frameworks.

Keywords: gender bias, artificial intelligence, gender equality, EU AI Act, AI governance

Introduction

The rapid proliferation of artificial intelligence systems across virtually every sector of society has brought unprecedented opportunities for innovation and efficiency, yet it has simultaneously exposed deep-rooted biases that reflect and amplify existing societal inequalities. Among the most pervasive and concerning of these biases is gender discrimination, which manifests across AI applications, from facial recognition systems that misidentify women at alarming rates to hiring algorithms that systematically favour male candidates. As AI systems increasingly influence crit-

© 2026 Zeynep Ayata published by Sciendo. This work is licensed under the Creative Commons Attribution-Non-Commercial-NoDerivatives 4.0 License.

ical decisions affecting employment, healthcare, criminal justice, and access to services, the embedded gender biases within these technologies pose significant threats to gender equality and women's fundamental rights.

These biases emerge not in a vacuum, but against a backdrop of profound gender disparities in the technology sector itself. Women remain dramatically underrepresented in artificial intelligence development and research, constituting only 12% of AI researchers globally and a mere 6% of software developers (Manasi et al., 2022). This underrepresentation extends throughout the technology pipeline: women earn only 18% of computer science bachelor's degrees in the United States, hold 26% of computing-related jobs, and occupy just 25% of technical roles at major technology companies. In academic settings, women comprise only 22% of AI professors globally, declining to 16% at full professor level, while representing merely 18% of presenters at leading AI conferences. The gender gap widens further when we examine leadership positions, with women holding only 15% of AI research director positions at major technology companies and 10% of leadership roles in AI start-ups. These structural inequalities in who develops AI systems directly shape the technologies produced, as homogeneous development teams are less likely to identify potential biases or consider diverse user needs and experiences (O'Connor & Liu, 2024).

The intersection of gender and artificial intelligence represents a complex sociotechnical challenge that extends far beyond mere algorithmic fairness. Gender bias in AI systems emerges through multiple pathways: biased training data that reflects historical patterns of discrimination, algorithmic design choices made by predominantly male development teams, and the broader cultural contexts that shape how these technologies are conceived, developed, and deployed. From virtual assistants programmed with submissive feminine personas to medical AI systems trained primarily on male patients, these biases are not incidental flaws but structural features that require systematic analysis and intervention.

This article examines the multifaceted nature of gender bias in AI systems through both theoretical and regulatory lenses, analysing how these biases manifest, persist, and can be addressed through comprehensive policy frameworks. The first section provides a thorough review of existing literature on gender bias in AI, exploring theoretical frameworks that explain how bias becomes embedded in algorithmic systems and examining empirical studies that document the scope and impact of gender discrimination across various AI applications. This analysis reveals how AI systems often perpetuate existing inequalities while creating new forms of digital discrimination that disproportionately affect women and marginalized gender groups. The second section shifts focus to policy and regulatory responses, examining how international organizations like UNESCO and the United Nations are developing frameworks to address gender equality in AI governance. Particular attention is given to the European Union's Artificial Intelligence Act, the world's first comprehensive AI regulation, which offers both opportunities and limitations for addressing gen-

der bias in AI systems. The analysis also considers the Council of Europe Framework Convention on Artificial Intelligence, the first legally binding international treaty in this domain, which establishes equality and non-discrimination as fundamental principles. Through critical analysis of these regulatory frameworks, this article evaluates whether existing approaches adequately address the complex challenges of gender bias in AI or whether more targeted interventions are necessary to ensure that AI serves as a tool for advancing rather than undermining gender equality in the digital age.

1. Theoretical frameworks and comprehensive analyses of gender bias in AI systems

Scholars in technology studies, such as Orlikowski and Fountain, argue that technologies are not neutral but rather reflect the social contexts in which they are created and deployed (Fountain, 2004; Orlikowski, 1992). Gender bias, defined as prejudiced actions based on the perception that women are not equal to men in rights and dignity, becomes embedded in AI systems through biased training data, algorithmic design choices, and the broader cultural contexts that shape technology development. This bias manifests both through language patterns and visual representations, with AI systems often amplifying existing societal inequalities rather than simply reflecting them. The literature review in this section examines various quantitative and qualitative analyses conducted in this field, addressing both general theories on how AI generates gender biases and specific problems in areas such as facial recognition, labour markets, and healthcare services.

A comprehensive framework for analysing gender bias has been presented by O'Connor and Liu (2024), who examine how AI systems perpetuate existing biases. The authors developed a two-dimensional analytical framework that categorizes AI technologies based on their data sources (text versus images) and their relationship to gender bias (perpetuation versus mitigation). For text-based AI, they examined how Google Translate perpetuates gender stereotypes by consistently translating gender-neutral pronouns as male for professional occupations, particularly in STEM fields. This bias was so pronounced that while women make up 35.94% of occupations according to Bureau of Labor Statistics data, they were represented with female pronouns in only 11.76% of translations. Conversely, O'Connor and Liu highlight successful bias mitigation efforts by researchers who developed debiasing algorithms for word embeddings, reducing gender stereotypes in semantic associations from 19% to 6%. In the realm of image-based AI, O'Connor and Liu (2024) analyse Joy Buolamwini and Timnit Gebru's (2018) groundbreaking 'Gender shades' study, which revealed significant intersectional biases in commercial facial recognition systems. These systems showed error rates of 8.1% to 20.6% between male and female classifi-

cations, with darker-skinned females experiencing misclassification rates as high as 72.4%. The study's impact extended beyond academic circles, prompting direct responses from major technology companies like IBM and Microsoft, who acknowledged the bias and committed to improving their systems.

Manasi et al. (2022) focus particularly on virtual assistants and robotics through the lens of feminist theory and the concept of affective labour. Their study provides detailed analysis of virtual assistants like Siri, Alexa, and Cortana, and notes that these systems are deliberately designed with feminine characteristics, including voices, names, and submissive personalities, that reinforce traditional gender stereotypes. The authors point out that while Google Assistant avoids gendered naming, most virtual assistants come with feminine voices and are programmed to perform traditionally female-coded tasks such as scheduling, note-taking, and list-making. This 'anthropomorphization' process creates what the authors term a 'master–servant relationship' similar to domestic labour arrangements. Problematically, these systems often fail to recognize or appropriately respond to gendered experiences, such as when early versions of Siri could not comprehend statements about sexual violence. In examining robotics, the authors highlight how service robots in sectors like hospitality, healthcare, and retail – traditionally seen as 'women's terrain' – are increasingly feminized and designed to perform affective labour. Research shows that 'male' robots are deemed appropriate for security-related jobs while 'female' robots are selected for healthcare settings, reflecting broader societal gender biases. The authors note that AI systems promise 'the allure of objectivity without public accountability' while embedding social biases (Manasi et al. 2022, p. 301). This is exacerbated by the severe underrepresentation of women in AI development – only 12% of AI researchers are women, and they represent just 6% of software developers.

Otis et al. (2024) present comprehensive evidence of a significant and persistent gender gap in generative AI use worldwide. The researchers synthesized data from 18 studies encompassing over 143,000 individuals across diverse regions, sectors, and occupations, combined with internet traffic data from major AI platforms like ChatGPT, Claude, and Perplexity. Their findings reveal a remarkably consistent pattern: women are approximately 20% less likely than men to use generative AI tools, with this gap holding across nearly all the contexts examined. According to the authors, the scale and universality of this gender disparity is striking. Analysis of representative US samples shows gaps of 10–20 percentage points, with similar patterns observed among populations ranging from science postdocs to business owners to college students across multiple countries. Internet traffic data corroborates these survey findings, showing that women comprise only 42% of ChatGPT website users globally and just 27% of mobile app downloads. To test whether the gap stems from differential access to the technology, the researchers conducted a novel experiment in Kenya, offering 17,541 participants equal opportunity to try ChatGPT. Even with access equalized, women remained 13% less likely to adopt the technology, indicating

that simply providing access is insufficient to close the gender divide. This suggests deeper underlying mechanisms driving the disparity, including differences in knowledge and familiarity with AI tools, confidence in using the technology effectively, and perceptions about the ethics of AI usage.

An empirical study by Ahn et al. (2022) investigates how gender stereotypes influence consumer evaluations of AI agents' recommendations, and specifically examined the interaction between an AI's 'gender' and the type of product (utilitarian versus hedonic). The researchers explored whether people apply the same gender stereotypes to AI agents that they use in human interactions, and how this affects trust in AI recommendations for different product categories. The study carried out two experiments, involving 180 and 120 participants respectively, testing interactions with chatbots and AI speakers using recorded human voices. The findings revealed significant gender stereotype effects on perceptions of AI personalities. 'Female' AI agents were perceived as significantly warmer than 'male' AI agents, supporting traditional gender stereotypes, and 'male' AI agents were perceived as more competent than 'female' AI agents. More importantly, the study found crucial interaction effects that depended on the product type. For utilitarian products, participants showed more positive attitudes and higher purchase intentions when they were recommended by 'male' AI agents. Conversely, for hedonic products, participants responded more favourably to recommendations from 'female' AI agents.

Domnich and Anbarjafari (2021) investigated gender bias in deep-learning models for facial expression recognition, contributing to the broader field of Responsible AI by examining fairness in emotion recognition systems. The researchers conducted a comprehensive analysis using six different neural network architectures, systematically dividing both training and testing data by gender to create 'regular' models (trained on all data), 'male' models (trained only on male data), and 'female' models (trained only on female data). The key findings revealed significant variations in gender bias across different neural network architectures. The study found that models generally performed better on emotions that aligned with traditional gender stereotypes, with recognition of surprise being more accurate for 'males' and emotions like sadness and being upset better recognized in 'females'. Interestingly, happiness recognition remained relatively consistent across genders. The analysis revealed that more biased neural networks consistently showed larger accuracy gaps between 'male' and 'female' test sets.

Andrews and Bucher (2022) examine how AI systems used in hiring processes can perpetuate gender discrimination. Their paper analyses three main AI technologies used in hiring that potentially discriminate against women: CV scanning, one-way video interviews, and video game assessments. These seemingly neutral technologies can embed gender bias because they are often trained on data from predominantly male workforces, causing them to favour traditionally masculine traits and communication styles. Amazon's failed hiring algorithm serves as a promi-

nent example: after 500 attempts, the company's engineers could not create an unbiased system because their algorithm, trained on predominantly male employee data, systematically rejected women's CVs. It penalized applications containing the word 'women' and rejected candidates from women-only colleges, while sometimes recommending unqualified male candidates. The authors explain how these technologies perpetuate discrimination through various mechanisms. CV-scanning algorithms may favour 'active' verbs like 'executed' and 'captured' more commonly used by men, while penalizing collaborative language like 'we' that women often use. One-way video interviews can discriminate based on speech patterns, facial expressions, and communication styles that differ between genders due to socialization.

Lau (2023) examines the critical issue of gender bias in artificial intelligence systems used in women's healthcare, analysing the problem from legal, technological, and feminist perspectives across legal frameworks in the United Kingdom and Europe. The core argument centres on how androcentricity in medicine – the male body serving as the standard template – has created systemic biases that are now amplified through AI technologies. When AI systems are trained on historical medical data that predominantly reflects male experiences and biology, they perpetuate and amplify existing inequalities. For instance, women experience adverse drug reactions twice as often as men because clinical trials have historically excluded women, leading to dosing protocols based on male physiology. Similarly, conditions like acute myocardial infarction are often misdiagnosed in women because their symptoms differ from the male-centred diagnostic criteria that AI systems are trained to recognize.

As this literature review demonstrates, without deliberate action to close the gender gap in AI development and usage, generative AI risks not only perpetuating existing gender inequalities but potentially widening them, limiting society's ability to benefit from the diverse perspectives and contributions women could bring to this transformative technology. The evidence thus demonstrates that gender bias in AI systems is a multifaceted problem requiring comprehensive solutions across technical, social, and policy dimensions.

2. Policy and regulatory perspectives

In an interview with the UN organization UN Women (2025), Zinnya del Villar, a leading expert on responsible AI, explores how AI technologies, while transformative, can perpetuate and amplify existing gender inequalities when trained on biased data. Del Villar explains that AI gender bias occurs when systems learn from data filled with stereotypes, leading them to reflect and reinforce discriminatory patterns in their decision-making processes. The real-world impacts are significant and far-reaching: in healthcare, AI systems may focus more on male symptoms, potentially leading to misdiagnoses for women; voice assistants that default to female

voices reinforce stereotypes about women being suited for service roles; and language models often associate certain professions with specific genders. The report highlights documented cases, such as Amazon's discontinued AI recruitment tool from 2018 that favoured male CVs, and image recognition systems that have struggled to accurately identify women, particularly women of colour, with serious implications for law enforcement and public safety.

2.1. UNESCO recommendations: Integrating gender equality into AI principles

UNESCO's Global Dialogue on Gender Equality and AI (2020) identifies critical gaps in how gender considerations are addressed in AI ethics frameworks. The report emphasizes that gender equality must be treated as more than an add-on to existing principles, requiring instead a fundamental transformation in how AI systems are developed, deployed, and governed. The recommendations call for moving beyond technical fixes to address systemic inequalities embedded in AI development processes.

The integration of gender equality into AI principles must begin with inclusive development processes that ensure meaningful participation by gender equality experts and women throughout the principles' formulation, interpretation, application, monitoring, and recalibration. This participation should occur at all levels – intergovernmental, sectoral, and institutional – and must be sustained throughout the entire lifecycle of AI governance frameworks. Meaningful involvement by gender experts and women is essential because their lived experiences and specialized knowledge enable the identification of potential biases and discriminatory impacts that might otherwise remain invisible to homogeneous development teams. Gender experts bring critical analytical frameworks for understanding how power dynamics operate within technological systems, while women's participation ensures that diverse perspectives inform the interpretation and application of AI principles. This inclusive approach helps prevent the reproduction of existing inequalities and creates pathways for AI systems to actively advance gender equality. The process requires deliberate and thoughtful consideration of when gender should be made explicit versus implicit, with careful attention to where these references are placed within frameworks to ensure accountability and meaningful implementation rather than tokenistic inclusion.

Gender equality should be established as a stand-alone principle rather than being subsumed under broader categories like bias or fairness. UNESCO emphasizes that gender encompasses much larger concerns than algorithmic bias alone, including women's empowerment, representation in leadership roles, access to education and training opportunities, and participation in decision-making processes (UNESCO, 2020). The principle should be positioned prominently within frameworks, with clear implementation pathways and monitoring mechanisms. Effective integration requires adopting whole-society, systems-based, and lifecycle approaches that

consider AI's broader social and structural implications. This means addressing gender equality not just in specific algorithms or datasets, but throughout the entire AI ecosystem – from initial research and development through deployment, use, and ongoing monitoring. The approach must recognize relationships and power dynamics between different actors, on local to global scales, and address the responsibilities of both private and public sectors. Understanding these power dynamics is crucial because AI development and deployment occur within existing structures of gender inequality, where certain voices and interests dominate while others are marginalized. Power operates through multiple channels: in determining research priorities and funding allocations, in shaping technical standards and best practices, in controlling access to data and computational resources, and in defining what constitutes 'successful' AI implementation. Recognizing these dynamics helps identify whose interests are served by particular AI systems and whose are overlooked or harmed, enabling more equitable distribution of AI's benefits and more effective mitigation of its risks. This systemic view acknowledges that AI operates within existing social structures and can either reinforce or challenge gender inequalities.

The recommendations outline a three-dimensional approach to addressing gender equality in AI. First, avoiding harm requires the proactive identification and mitigation of AI's negative impacts on women and girls, including bias in algorithms, discriminatory outcomes, and reinforcement of harmful stereotypes. Second, increasing visibility involves ensuring that women's experiences, perspectives, and needs are represented in AI development and that gender implications are explicitly considered rather than overlooked. Third, contributing to empowerment means leveraging AI's potential to actively advance gender equality, challenge oppressive norms, and create opportunities for women's advancement in areas like education, economic participation, and political representation. This three-dimensional framework recognizes that gender equality in AI requires more than preventing discrimination – it demands proactive measures to empower women and transform existing power structures. By framing gender equality across these three dimensions, UNESCO emphasizes that AI governance must simultaneously protect against harms, ensure the representation and visibility of women's concerns, and actively promote women's empowerment and advancement. This comprehensive approach contrasts sharply with regulatory frameworks that focus primarily on harm prevention while neglecting the transformative potential of AI as a tool for advancing gender equality.

The UNESCO framework emphasizes the critical importance of intersectionality, recognizing that women's experiences vary significantly based on race, ethnicity, age, disability status, sexual orientation, geographic location, and socioeconomic status (UNESCO, 2020). AI principles must account for these multiple and overlapping forms of discrimination and ensure that solutions do not inadvertently harm marginalized groups while helping others. This requires diverse and inclusive teams in AI development that include not only women but also gender equality experts, repre-

sentatives from affected communities, and specialists in relevant domains where AI systems will be deployed.

Moving from principles to practice requires concrete mechanisms for implementation across multiple stakeholder groups. For the private sector, this includes establishing corporate governance mechanisms that integrate gender equality considerations, implementing bias detection and mitigation tools throughout the AI development lifecycle, creating diverse development teams with meaningful gender equality expertise, and conducting regular algorithmic impact assessments with gender-specific criteria. Government action should focus on developing appropriate policy frameworks, funding gender-responsive AI research and development, ensuring that procurement practices promote gender equality, and creating accountability mechanisms for AI systems used in public services.

The recommendations emphasize the need for robust monitoring and accountability mechanisms to ensure that gender equality commitments translate into measurable outcomes. This includes developing gender-responsive indicators for AI systems, establishing independent oversight bodies with gender equality expertise, creating transparent reporting requirements for AI developers and deployers, and implementing feedback mechanisms that centre the voices of affected women and marginalized communities. Regular assessment and recalibration of both principles and implementation strategies are essential to address emerging challenges and opportunities as AI technologies continue to evolve.

2.2. A general overview of the EU AI Act and the Council of Europe Framework Convention

The first-ever and most significant comprehensive regulation of AI is Regulation (EU) 2024/1689, known as the AI Act, adopted by the EU in 2024. The Act is therefore worth analysing from the point of view of gender policy and equality, as it might also constitute an inspiration or benchmark for regulation in other jurisdictions. Additionally, the Council of Europe Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law, adopted in May 2024 and opened for signature in September 2024, represents the first legally binding international treaty on AI. The Convention establishes human rights, democracy, and the rule of law as fundamental principles for AI systems, with equality and non-discrimination prominently featured among its core protections. The EU's participation in developing this Convention provides important context for understanding the Union's broader approach to AI regulation and gender equality. Although the AI Act will not be fully applicable until August 2026, it is important to examine the possible consequences of its implementation in terms of gender bias. This section will thus first account for the goals and general rules of the EU AI Act and will reference the Council of Europe Framework Convention, before examining in the subsequent section the specific treatment of gender equality within these frameworks.

The Artificial Intelligence Act, officially designated as Regulation (EU) 2024/1689 and enacted on 13 June 2024, represents the world's first comprehensive regulatory framework for artificial intelligence systems. It aims to improve the functioning of the internal market while promoting human-centric and trustworthy AI development, ensuring high levels of protection for health, safety, fundamental rights, democracy, the rule of law, and the environment against potential harmful effects of AI systems within the Union. The legislation adopts a risk-based approach to AI regulation, categorizing AI systems into different risk levels with corresponding obligations. At the most restrictive level, the Act prohibits certain AI practices deemed unacceptable, including AI systems that deploy subliminal techniques to materially distort human behaviour, exploit vulnerabilities of specific groups, implement social scoring systems leading to detrimental treatment, and create or expand facial recognition databases through untargeted scraping. The Act also heavily restricts real-time remote biometric identification systems in publicly accessible spaces for law enforcement purposes, permitting their use only in narrowly defined circumstances such as searching for victims of serious crimes or preventing terrorist attacks.

For high-risk AI systems, which include those used in critical infrastructure, education, employment, essential services, law enforcement, migration control, and democratic processes, the Act establishes mandatory requirements covering risk management, data governance, technical documentation, transparency, human oversight, and cybersecurity. These systems must undergo conformity assessments before market placement and bear a CE marking to indicate compliance. Providers must implement quality management systems, maintain detailed technical documentation, and establish post-market monitoring methods to track system performance throughout its lifecycle.

The Act introduces specific provisions for general-purpose AI models, particularly those with systemic risks identified by computational thresholds, such as models trained with more than 10^{25} floating point operations. Providers of these models must conduct model evaluations, assess and mitigate systemic risks, report serious incidents, and ensure adequate cybersecurity protection. The legislation encourages the development of codes of practice to demonstrate compliance and establishes the AI Office within the European Commission to monitor and enforce obligations for general-purpose AI models.

The governance structure includes the European Artificial Intelligence Board, composed of Member State representatives, a scientific panel of independent experts, and an advisory forum representing various stakeholders. Member States must designate national competent authorities for market surveillance and conformity assessment, with specific provisions for different sectors, including financial services and law enforcement. The Act provides for significant penalties, with administrative fines reaching up to EUR 35 million or 7% of worldwide annual turnover for prohibited AI

practices, and establishes procedures for market surveillance, incident reporting, and enforcement coordination across the Union.

The regulation will be implemented in phases, with prohibitions and general provisions applying from 2 February 2025, governance structures operational by 2 August 2025, obligations for general-purpose AI models effective from 2 August 2025, and the full regulation applying from 2 August 2026. This staggered implementation allows time for the development of the technical standards, codes of practice, and institutional frameworks necessary for effective enforcement, while providing legal certainty for AI developers and deployers across the European Union.

2.3. A critical assessment of the EU AI Act from a gender equality perspective

In order to analyse the Act from a gender equality perspective, we can start by examining its impact assessment report. It should first be noted that the report deals with and refers to gender equality and gender bias within the broad context of discrimination and not as a specific or separate problem. As we will see, this approach is reflected in the Act itself, where gender equality is not identified as a stand-alone target – a treatment that stands in stark contrast to UNESCO's recommendations and raises questions about whether this subsumption under general non-discrimination provisions adequately addresses the specific and multifaceted challenges of gender bias in AI. This pattern also reflects the EU's broader contemporary approach to gender equality in recent legal initiatives, where gender concerns are frequently integrated within general equality frameworks rather than being addressed through targeted mechanisms. The report addresses gender biases in the 'Algorithmic discrimination' section with reference to facial recognition systems, acknowledging that biases that occur for women may not occur for men due to data that 'might be unrepresentative, incomplete or contain historical biases'. According to the report, this concern is part of the general problem of biases in the existing data that AI depends on. The report acknowledges that use of discriminatory AI systems may lead to serious societal consequences whereby AI regenerates 'existing or create[es] new forms of structural discrimination and exclusion'.

The report also refers to gender equality concerns in the section called 'Requirements for trustworthy AI envisaged in the EU voluntary labelling scheme' (p. 41), where it explains certain policy decisions made in the AI Act. Accordingly, gender equality was one of the five requirements proposed by the European Parliament for high-risk systems. The Parliament's proposals during the legislative process included specific requirements addressing gender equality in high-risk AI systems, reflecting parliamentary concerns about the potential for AI to perpetuate gender discrimination. These proposals emerged from the Parliament's amendments to the Commission's original draft and sought to establish gender equality as an explicit requirement alongside other fundamental rights protections. The decision not to incorporate these proposals into the final Act represents a significant gap between parliamen-

tary advocacy for gender-responsive AI regulation and the adopted legislative framework. As stated explicitly in the report, 'it was decided not to include some of the [European Parliament] proposals or the High Level Expert Group on AI's principles as requirements […] because they were considered […] too vague for a legal act and too difficult to operationalize'. Certain other requirements proposed by the Parliament were not included on the grounds that they were already covered by other EU legal acts, such as the example of privacy and the GDPR. The report does not make such a claim for gender equality, implicitly admitting that there are no existing legal rules that would comprehensively cover this area in the AI context. The EU has also adopted the Directive on Violence against Women and Domestic Violence, which specifically criminalizes the non-consensual production of AI-generated or manipulated material depicting persons in sexually explicit activities (deepfakes), showing the EU's broader commitment to addressing gender-related AI harms. However, the impact assessment report does not refer to this Directive as legal grounds for ensuring gender equality in the context of AI systems. Instead, the report claims that this requirement would not be specific enough or would be too difficult to implement.

The AI Act itself contains few direct references to gender equality in the main text. While it refers to vulnerable persons or groups seven times, gender is not conceived as a factor of vulnerability. In the recitals, gender-related references are more frequent but still limited. While 'fundamental rights' appears 61 times and 'discrimination' 33 times, 'women' and 'gender' appear only two and four times respectively. Recital 27, which outlines the ethical principles for trustworthy AI, states that '[d]iversity, non-discrimination and fairness means that AI systems are developed and used in a way that includes diverse actors and promotes equal access, gender equality and cultural diversity, while avoiding discriminatory impacts and unfair biases'. Recital 48 refers to gender equality as a fundamental right protected under the EU Charter of Fundamental Rights, underlining that any violation of such a right would play an important role in classifying an AI system as high-risk. Recital 58 draws attention to the specific problem where AI systems are used to determine potential beneficiaries of public or private services and particularly the right to certain financial resources. It acknowledges that AI systems used for such purposes may lead to discriminatory results or 'may perpetuate historical patterns of discrimination', including gender biases; such consequences will be considered when assessing the risk level of AI systems. On the other hand, the term '(non-)discrimination' appears 16 times throughout the Act, and 'women' is mentioned twice, indicating that gender equality is primarily addressed within the broader non-discrimination framework rather than as a stand-alone concern.

In the main articles, the term 'fundamental rights' appears 45 times, while 'bias' or 'biases' appears nine times and 'discrimination' only twice. The term 'gender' is recorded just twice, with only one explicit reference to gender equality in Article 95, which deals with voluntary codes of conduct for non-high-risk AI systems. Al-

though the Act does not establish a specific target for gender equality or provide special prohibitions for gender biases directly, it includes provisions for bias audits and fundamental rights impact assessments that can help reduce gender biases and discriminatory risks in AI systems. However, critics note that while the Act aims to safeguard fundamental rights and address gender discrimination, it lacks comprehensive mechanisms specifically for gender-related issues and does not mandate fundamental rights impact assessments for all systems. It is therefore worth analysing various articles of the Act that may have indirect consequences for gender equality and examining how they have been received in academic discussions.

Article 5 of the AI Act establishes prohibitions on AI practices deemed unacceptable due to their potential to jeopardize safety and fundamental rights, explicitly banning systems such as social scoring and manipulative techniques targeting vulnerable individuals. While the provision prohibits biometric categorization systems in public spaces involving sensitive characteristics such as race and sexual orientation, it notably omits gender from this protected list, raising significant concerns about gender protection in contexts of social behaviour and performativity. In a comprehensive report penned for the Friedrich Ebert Stiftung, Karagianni (2025b) identifies a critical oversight in this approach to vulnerability in Article 5 of the Act, drawing attention to the concept of vulnerability as understood in both computer science and social sciences and highlighting how it particularly affects marginalized groups. Despite recommendations from the European Data Protection Board to include gender as a sensitive category, this has not been integrated into the AI Act. Furthermore, as indicated above, the UNESCO recommendations also emphasize the importance of making gender equality a specific target within vulnerable groups (UNESCO, 2020). This omission by the AI Act represents a fundamental gap in ensuring gender equality in AI systems and underscores the need for an intersectional approach, as outlined by UNESCO, that acknowledges the complex layers of discrimination inherent in AI technologies. The failure to recognize gender as a sensitive characteristic leaves women and gender minorities vulnerable to discriminatory treatment by AI systems operating in public spaces.

Data governance is one of the most critical areas for addressing gender equality concerns. The data governance requirements in the AI Act's Article 10 will be operationalized through the future European standard on 'Data and data governance', which aims to develop specifications for adequate data governance and data management procedures for AI system providers. The standard has a twofold purpose: first, to establish specifications for data governance and management procedures focusing on data generation, collection, preparation operations, design choices, and procedures for detecting and addressing biases; and second, to provide specifications on quality aspects of the datasets used to train, validate, and test AI systems, including requirements for representativeness, relevance, completeness, and correctness. The 'gender data gap' within this framework is crucial: datasets often suffer from being

non-representative, incomplete, and incorrect due to the digital gender divide and prevailing gender stereotypes. Lütz (2024) argues that this standard would benefit significantly from incorporating gender equality expertise and ensuring women's involvement in both its development and implementation phases to enable diverse perspectives and identify potential pitfalls in data governance practices. According to Lütz, the doctrine and institutional reports have clearly identified design choices and datasets as potential entry points for gender biases and discrimination throughout the algorithmic development process, from data collection and generation to the modification and preparation of training datasets. Given that bias detection and subsequent addressing of gender biases serve as crucial tools for achieving gender equality, clear guidance is essential for both companies and enforcement authorities.

Article 27 outlines the Fundamental Rights Impact Assessment (FRIA) as a solution for protecting fundamental rights endangered by high-risk AI systems, working alongside the risk management system detailed in Article 9. However, questions arise regarding their adequacy in preventing gender-based discrimination and promoting gender equality within AI contexts. While these represent novel measures in AI regulation, risk management and impact assessments are not new constructs in technology regulation, having historically emerged to address uncertainties associated with technological advancements.

The existing literature lacks comprehensive analysis of the differences between and practical applications of risk management systems and FRIAs, particularly in AI contexts. Defining what constitutes a risk to rights such as non-discrimination involves multiple conceptualizations, complicating risk measurement as a subjective process often influenced by gendered assumptions. The relationship between risk management systems, FRIAs, and gender impact assessments, which specifically address impacts on gender equality, remains unclear and requires clarification. Karagianni argues that incorporating gender impact assessments into risk management systems and fundamental rights impact assessments is essential, reinforcing the principle that 'women's rights are human rights' (Karagianni, 2025b). Without this integration, these assessment tools risk protecting only the rights of a standard, liberal legal persona, typically male, white, and able-bodied. This claim is also parallel to the recommendations of UNESCO on the responsibilities of governments to establish adequate monitoring and assessment mechanisms. Gender impact assessments highlight how development initiatives affect individuals differently based on their gender, and aim to uncover disparities caused by entrenched structural inequalities. Examples include the identification of potential gender-based impacts, assessment of whether AI systems reinforce existing inequalities, examination of intersectional factors, consideration of gender-specific rights under international treaties, and collection of gender-disaggregated data to understand different experiences and needs.

For general-purpose AI systems under Article 51 and post-release obligations under Article 26, while the Act mandates ongoing monitoring to detect evolv-

ing biases, it relies heavily on self-regulation and lacks robust redress mechanisms (Karagianni, 2025a). This is particularly problematic because gender bias in AI disproportionately harms marginalized communities who often lack institutional power to demand accountability. The Act provides no clear pathways for individuals affected by algorithmic discrimination to seek justice, such as women rejected by biased AI hiring systems.

The standardization process established in Articles 40 and 41, while creating important technical standards through bodies like European Committee for Standardization (CEN), European Committee for Electrotechnical Standardization (CENELEC) and European Telecommunications Standards Institute (ETSI), is criticized for lacking mandatory intersectional gender audits and diverse AI development teams (Lütz, 2024). The AI standardization process outlined in Article 40 is essential for establishing comprehensive guidelines, technical specifications, and best practices that ensure AI systems are safe, reliable, transparent, and ethically designed. This standardization involves the development and implementation of norms and technical standards by various entities, including European standards organizations, national standards bodies, and the European Commission, working together to establish frameworks regulating AI with a focus on human-centred AI, security, privacy, and data governance. The standardization process faces significant challenges regarding inclusive participation. While the European Commission has emphasized the need to include representatives from various sectors and organizations, achieving consensus on important issues has proven difficult. Karagianni (2025b) emphasizes the importance of explainability and accountability within AI systems as critical to upholding fundamental rights. There is also substantial risk of tokenism, where the involvement of diverse stakeholders remains superficial and fails to result in substantive changes promoting gender equity. This challenge is exacerbated by the AI standardization process being predominantly shaped by powerful corporations and governments, which may obstruct genuine feminist efforts to cultivate inclusivity and equity in AI development and regulation.

High-risk AI systems are mandated to undergo conformity assessments during their evaluation stage to ensure adherence to established safety and ethical standards before market release. This process verifies system compliance with necessary requirements and creates a comprehensive regulatory framework for such systems. Article 43 works in conjunction with Article 6 to establish governance structures aimed at overseeing compliance and enforcement, promoting responsible AI development while safeguarding individual rights and interests.

From a feminist perspective, AI conformity assessment must address gender bias, inclusivity, and power structures, particularly given systemic inequalities that may be perpetuated if assessments lack an intersectional lens. A prominent feminist concern is the reinforcement of gender bias through biased algorithms and data, as AI systems often rely on historical data reflecting existing social inequalities. Consequently,

conformity assessments should require comprehensive bias audits extending beyond identifying overt discrimination to pinpointing subtle structural biases against women, particularly women of colour, LGBTQIA+ individuals, and other marginalized groups. Intersectional data analysis in conformity assessments is essential, as gender bias should not be evaluated in isolation (UNESCO, 2020). Assessments must consider how gender intersects with race, class, disability, and other identity factors to ensure AI systems do not disproportionately harm marginalized communities. The assessment process should involve diverse teams reflecting a range of genders, races, and social backgrounds to ensure AI systems are scrutinized through multiple lenses, minimizing bias and enhancing fairness (Karagianni, 2025b).

The Council of Europe Framework Convention on Artificial Intelligence provides an important comparative perspective on how gender equality can be integrated into international AI governance. As the first legally binding international treaty on AI, the Convention establishes equality and non-discrimination as fundamental principles that must be respected throughout the lifecycle of AI systems. The Convention's approach has prompted scholarly analysis of its implications for gender equality, with researchers like Bartoletti and Xenidis (2023) examining how its provisions address discrimination and equality concerns. The EU's participation in developing this Convention reflects the broader European commitment to human rights-based AI governance, yet the relationship between the Convention's equality provisions and the AI Act's treatment of gender remains an area requiring further harmonization and clarity.

The policy and regulatory landscape reveals a fundamental tension between aspirational frameworks and practical implementation in addressing gender bias in AI systems. While UNESCO's recommendations provide a comprehensive blueprint for integrating gender equality as a central principle in AI governance, emphasizing meaningful participation, intersectionality, and transformative approaches that go beyond harm prevention, the EU AI Act's treatment of gender issues remains fragmented and inadequate. Despite being the world's first comprehensive AI regulation, the Act's failure to establish gender equality as a stand-alone principle, its omission of gender from protected characteristics in key provisions, and its reliance on general non-discrimination frameworks rather than targeted gender-specific mechanisms highlight the significant gap between recognition of the problem and regulatory solutions. This treatment reflects a broader pattern in the EU's recent legal initiatives, where gender equality concerns are frequently subsumed within general equality and diversity frameworks rather than receiving dedicated attention. The EU's exclusive focus on avoiding harm, to the exclusion of UNESCO's other two dimensions – increasing visibility and contributing to empowerment – further limits the transformative potential of AI regulation for advancing gender equality. This disparity underscores the urgent need for more deliberate and comprehensive integration of gender equality considerations into AI governance frameworks, moving beyond to-

kenistic inclusion toward substantive transformation of how AI systems are developed, deployed, and monitored.

Conclusions

The analysis of gender bias in AI systems and regulatory responses reveals a significant disconnect between the comprehensive frameworks proposed by international organizations and the practical implementation found in existing legislation. While the UNESCO recommendations provide a holistic, transformative approach to integrating gender equality into AI governance, the European Union's AI Act, despite being groundbreaking in its scope, falls short of adequately addressing the complex challenges of gender discrimination in artificial intelligence systems.

The UNESCO framework's emphasis on treating gender equality as a stand-alone principle rather than subsuming it under broader categories of bias or discrimination stands in stark contrast to the EU AI Act's approach. Where UNESCO calls for gender equality to be positioned prominently within frameworks with clear implementation pathways, the EU Act relegates gender considerations to occasional mentions within broader non-discrimination provisions. This fundamental difference reflects deeper philosophical divides about whether gender bias requires specialized attention or can be adequately addressed through general fairness mechanisms. The EU's treatment of gender equality within general non-discrimination frameworks mirrors a broader tendency in the Union's recent legal initiatives, suggesting systemic challenges in recognizing and addressing the specific dimensions of gender inequality in technological contexts.

Most critically, the UNESCO recommendations advocate for a three-dimensional approach – avoiding harm, increasing visibility, and contributing to empowerment – that goes far beyond the EU Act's focus primarily on harm prevention. The EU's exclusive focus on avoiding harm, without incorporating mechanisms for increasing women's visibility in AI development or leveraging AI's potential for empowerment, represents a missed opportunity for transformative change. While the EU Act establishes important prohibitions and risk management requirements, it lacks the proactive mechanisms necessary to leverage AI's potential for advancing gender equality. The absence of mandatory gender impact assessments, the omission of gender from sensitive characteristics in biometric categorization prohibitions, and the limited pathways for redress all highlight the Act's reactive rather than transformative approach to gender equality. This narrow focus fails to address why gender equality was excluded from the final regulatory framework despite being proposed by the European Parliament, suggesting that concerns about operationalization and vagueness may have overshadowed the fundamental importance of targeted gender equality provisions.

The standardization processes outlined in the EU Act present both opportunities and risks for gender equality. While these processes could potentially incorporate gender expertise and diverse perspectives, the current framework relies heavily on self-regulation and lacks mandatory requirements for intersectional gender audits or diverse development teams. This contrasts sharply with UNESCO's emphasis on the meaningful participation of gender equality experts throughout the entire lifecycle of AI governance frameworks. The meaningful participation that UNESCO envisions – involving gender experts and women in formulation, interpretation, application, and monitoring – requires institutional mechanisms that recognize and value their contributions, ensure their voices shape decisions rather than merely being consulted, and create pathways for their insights to influence technical standards and implementation practices. The conformity assessment mechanisms in the EU Act, while innovative, require significant enhancement to address structural gender biases effectively. The current assessment framework lacks the intersectional lens that UNESCO identifies as essential for understanding how gender intersects with race, class, disability, and other identity factors. Without comprehensive bias audits that extend beyond identifying overt discrimination to uncover subtle structural biases, these assessments risk perpetuating existing inequalities while providing a veneer of compliance. The Council of Europe Framework Convention on Artificial Intelligence offers an important complementary perspective, establishing equality and non-discrimination as fundamental principles in the first legally binding international treaty on AI. However, the relationship between the Convention's approach and the EU Act's provisions requires further examination to ensure the coherent and comprehensive protection of gender equality across European AI governance frameworks.

Moving forward, the implementation of the EU AI Act presents a critical opportunity to bridge the gap between aspirational principles and regulatory practice. As the Act's provisions come into force through 2026, there is an urgent need to incorporate gender equality expertise into the development of technical standards, ensure the meaningful participation of diverse stakeholders in governance structures, and establish robust monitoring mechanisms that centre the voices of affected women and marginalized communities. The implementation phase must address why the European Parliament's proposals for explicit gender equality requirements were deemed too vague or difficult to operationalize, and explore concrete mechanisms for translating gender equality principles into enforceable standards. Only through such comprehensive integration can AI regulation move beyond protecting the status quo to actively promoting gender equality in the digital age. The stakes are too high, and the potential for both harm and advancement too great, to accept anything less than a truly transformative approach to gender equality in AI governance.

REFERENCES

Ahn, J., Kim, J., & Sung, Y. (2022). The effect of gender stereotypes on artificial intelligence recommendations. *Journal of Business Research*, *141*, 50–59.

Andrews, L., & Bucher, H. (2022). Automating discrimination: AI hiring practices and gender inequality. *Cardozo Law Review*, *44*, 145–178.

Bartoletti, I., & Xenidis, R. (2023). The Council of Europe's Framework Convention on Artificial Intelligence: Equality and non-discrimination perspectives. *European Equality Law Review*, *1*, 56–72.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability and Transparency*, *81*, 77–91.

Domnich, A., & Anbarjafari, G. (2021). Responsible AI: Gender bias assessment in emotion recognition. *arXiv preprint*. arXiv:2103.11436.

Fountain, J. E. (2004). *Building the virtual state: Information technology and institutional change*. Brookings Institution Press.

Karagianni, A. (2025a). Gender in a stereo-(gender)typical EU AI law: A feminist reading of the AI Act. *Cambridge Forum on AI: Law and Governance*, *1*(e25), 1–18.

Karagianni, A. (2025b). *The EU Artificial Intelligence Act through a gender lens*. Friedrich-Ebert-Stiftung e.V. https://library.fes.de/pdf-files/bueros/bruessel/21887–20250304.pdf

Lau, P. L. (2023). AI gender biases in women's healthcare: Perspectives from the United Kingdom and the European legal space. In E. Gill-Pedro & A. Moberg (Eds.), *YSEC yearbook of socio-economic constitutions 2023: Law and the governance of artificial intelligence* (pp. 247–274).

Lütz, F. (2024). The AI Act, gender equality and non-discrimination: What role for the AI office? *ERA Forum*, *25*, 79–95.

Manasi, A., Panchanadeswaran, S., Sours, E., & Lee, S. J. (2022). Mirroring the bias: Gender and artificial intelligence. *Gender, Technology and Development*, *26*(3), 295–305.

O'Connor, S., & Liu, H. (2024). Gender bias perpetuation and mitigation in AI technologies: Challenges and opportunities. *AI & Society*, *39*, 2045–2057.

Orlikowski, W. J. (1992). The duality of technology: Rethinking the concept of technology in organizations. *Organization Science*, *3*(3), 398–427.

Otis, N. G., Delecourt, S., Cranney, K., & Koning, R. (2024). *Global evidence on gender gaps and generative AI* [Working paper 25–023]. Harvard Business School.

UN Women. (2025, 5 February). *How AI reinforces gender bias – and what we can do about it: Interview with Zinnya del Villar on AI gender bias and creating inclusive technology*. https://www.unwomen.org/en/news-stories/interview/2025/02/how-ai-reinforces-gender-bias-and-what-we-can-do-about-it

UNESCO. (2020). *Artificial intelligence and gender equality: Key findings of UNESCO's global dialogue*. https://unesdoc.unesco.org/ark:/48223/pf0000374174/PDF/374174eng.pdf.multi