STUDY

Requested by the FEMM committee

European Parliament

# Gender bias and artificial intelligence



Policy Department for Citizens, Equality and Culture
Authors: Nele ROEKENS, Carlotta RIGOTTI
Directorate-General for Citizens' Rights, Justice and Institutional Affairs
PE 791.844 – November 2026

EN

# Gender bias and artificial intelligence

**Abstract**

This study analyses gender bias in AI and examines emerging forms of technology-facilitated gender-based violence through an intersectional lens, examining its manifestations and impacts on women across domains. The study identifies legal, governance and technical responses, and assesses their effectiveness in preventing and providing redress for algorithmic gender discrimination and technology-facilitated gender-based violence.

This document was provided by Policy Department Citizens, Equality and Culture at the request of the FEMM Committee.

This document was requested by the European Parliament's Committee on Women's Rights and Gender Equality.

**AUTHOR(S)**

Nele Roekens, LL.M., Legal Advisor, Equinet

Dr. Carlotta Rigotti, LL.M., Assistant Professor, Leiden University

**ACKNOWLEDGEMENTS**

The authors would like to thank Ruurd Zanting, Nathalie Smuha and Kris Shrishak for their review contributions, and to Laura Lázaro Cabrera and Simona De Heer for their expert input. We are grateful to the Equinet secretariat, and Marta Pompili, Milla Vidina and Mariam Rechchad in particular, for their input and coordinating outreach to their members. Lastly, we are grateful to all the participating stakeholders for their valuable input and recommendations. Nele would also like to specifically thank Zadie.

**LINGUISTIC EDITOR**

Rebecca CHIRCOP

**CONTACTS IN THE EUROPEAN PARLIAMENT**

Coordination: Georgiana SANDU
Editorial assistance: Dana AXANEI
To give feedback or obtain copies, please write to: poldep-iust-a@europarl.europa.eu

**PEER REVIEW**

This study was subject to an external peer review.

**VERSION(S)**

Original: English
Translation: N/A
Manuscript completed: November 2026.

**BIBLIOGRAPHIC REFERENCE FOR THIS PAPER**

ROEKENS, N. and RIGOTTI C., *Gender bias and artificial intelligence*, Brussels, European Parliament, Policy Department for Citizens, Equality and Culture.

For in-text citations: Roekens and Rigotti, 2026

**DISCLAIMER**

The opinions expressed in this publication are those of the authors only and should not be considered as representative of the European Parliament's official position.

**ARTIFICIAL INTELLIGENCE**

Nele Roekens used licensed version of Lumo, an AI model developed by Proton AG, and a licensed version Co-Pilot, a model developed by Microsoft, both for language editing, broadening the range of sources available to the author and assisting in ensuring the proper format for bibliographic references. All analysis and conclusions are the author's own, and the author remains responsible for the content of this study.

Carlotta Rigotti used Claude, an AI model developed by Anthropic, for language editing. Claude was also used to translate documentation published in languages the author does not speak, where these sources had first been identified through references in English-language media and other publications. All analysis and conclusions are the author's own, and the author remains responsible for the content of this study.

Any AI-generated content in this text has been reviewed by the authors.

Rebecca Chircop used Claude, an AI model developed by Anthropic, to check for internal stylistic inconsistencies within the document, and to assist with its review by, for example, verifying consistent adherence with the style guide's principles following the initial review.


## COPYRIGHT

© European Union, 2026
Image on cover: Adobe Stock

## LICENCE

The reuse of this document is authorised under a Creative Commons Attribution 4.0 International (CC-BY 4.0) licence (https://creativecommons.org/licenses/by/4.0/). To use or reproduce elements that are not owned by the European Union, permission may need to be sought directly from the respective rightsholders.

# CONTENTS


**LIST OF ABBREVIATIONS** 6

**LIST OF BOXES** 9

**LIST OF TABLES** 9

**EXECUTIVE SUMMARY** 10

**1. INTRODUCTION** 13

1.1. Research questions 13

1.2. Methodology 14

1.3. Structure of the study 14

1.4. Scope 14

**2. AI AND GENDER BIAS: CONCEPTS, DEFINITIONS AND SOURCES** 16

2.1. The concept of artificial intelligence 16

2.2. The definition of AI at the EU level 17

2.3. The concept of gender bias in AI 18

2.4. The sources of gender bias in AI 20

2.4.1. Pre-existing bias 21

2.4.2. Technical bias 21

2.4.3. Emergent bias 23

2.4.4. Concluding considerations 24

**3. HOW GENDER BIAS IN AI AFFECTS WOMEN: EVIDENCE FROM FOUR DOMAINS** 25

3.1. Employment and the labour market 26

3.1.1. Exposure to AI and equal access 26

3.1.2. Gender bias from recruitment to workplace 27

3.1.3. Intersecting inequalities in the automated workplace 28

3.2. Public and private services 28

3.2.1. Welfare and social security 28

3.2.2. Credit scoring, lending, and insurance 29

3.3. Healthcare 30

3.3.1. Gender bias in clinical and administrative uses of AI 30

3.3.2. The case of femtech 31

3.4. Social media 32

3.4.1. Recommender systems 32

3.4.2. Content moderation 33

3.4.3. Advertising systems 33
3.5. Other domains 34
**4. TECHNOLOGY-FACILITATED GENDER-BASED VIOLENCE: A PERVASIVE AND EVOLVING PHENOMENON 35**
4.1. Terminology 35
4.2. Manifestations 37
4.2.1. Image-based sexual abuse 37
4.2.2. Sexual digital forgeries 38
4.2.3. Sextortion 39
4.2.4. Cyberharassment 40
4.2.5. Cyberstalking 40
4.2.6. Gender-based hate speech 41
4.2.7. Meta-rape 42
4.2.8. Chatbot-simulated and sex robot-mediated abuse 42
4.3. Nature and impact 42
4.4. Scale and data collection 44
**5. LEGAL FRAMEWORK 46**
5.1. The anti-discrimination framework 47
5.1.1. The scope of EU gender equality and anti-discrimination law 47
5.1.2. Prevailing forms of algorithmic discrimination 48
5.2. The European Union's digital acquis 51
5.2.1. The Artificial Intelligence Act 51
5.2.2. The General Data Protection Regulation 55
5.2.3. The Digital Services Act 56
5.2.4. The Directive on combatting violence against women and domestic violence 56
5.2.5. Recent EU legislative and regulatory developments 57
5.2.6. The Council of Europe's Framework 59
5.3. Implementation challenges 59
**6. MAPPING OF TECHNICAL, ORGANISATIONAL AND GOVERNANCE MEASURES TO PREVENT AND MITIGATE GENDER BIAS 61**
6.1. AI Act requirements 62
6.2. Socio-technical measures to prevent and mitigate gender bias 63
6.2.1. Bias mitigation measures 63
6.2.2. Fairness metrics 68
6.3. Organisational and governance measures to prevent and mitigate gender bias 70

6.3.1. Fundamental rights and equality impact assessments, participation and prevention 70
6.3.2. The use of AI to advance gender equality 72
6.4. Selection of best practices at European and national level 73
**7. REDRESS, OVERSIGHT AND ACCOUNTABILITY MECHANISMS AND THEIR EFFECTIVENESS 75**
7.1. Barriers to effective redress for AI-related harm 76
7.1.1. Barriers relating to the identification and reporting of harm 76
7.1.2. Barriers to redress arising from the intersectional, systemic, collective and societal nature of algorithmic harm 77
7.1.3. Resource and expertise barriers 78
7.1.4. Institutional and procedural barriers, including transnational dimension 79
7.2. Emerging rights and mechanisms addressing technology-specific barriers to redress 80
7.2.1. Information and transparency rights 80
7.2.2. Complaint and individual redress mechanisms 82
7.2.3. Collective redress mechanisms 83
7.2.4. Oversight and enforcement through and in cooperation with specialised actors 84
7.2.5. Accountability through self-assessment and risk management 85
7.2.6. Liability, causation and evidentiary burdens 85
**8. RECOMMENDATIONS 88**
**REFERENCES 91**
**ANNEX 1: METHODOLOGICAL APPROACH 122**
**ANNEX 2: CONSULTATION QUESTIONNAIRE ON GENDER BIAS AND ARTIFICIAL INTELLIGENCE 124**
**ANNEX 3: SIMPLIFIED VERSION OF THE QUESTIONNAIRE 126**

# LIST OF ABBREVIATIONS

| | |
|---|---|
| **ADM** | Automated Decision-Making |
| **AI** | Artificial Intelligence |
| **AI Act** | Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence |
| **AI HLEG** | High-Level Expert Group on Artificial Intelligence |
| **AILD** | Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive) |
| **CCTV** | Closed-circuit television |
| **CEDAW** | United Nations Convention on the Elimination of All Forms of Discrimination against Women |
| **CEN** | European Committee for Standardization |
| **CENELEC** | European Committee for Electrotechnical Standardization |
| **CSOs** | Civil Society Organisations |
| **CJEU** | Court of Justice of the European Union |
| **COE** | Council of Europe |
| **DSA** | Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) |
| **ECHR** | European Convention on Human Rights |
| **ECtHR** | European Court of Human Rights |
| **EIGE** | European Institute for Gender Equality |
| **EU** | European Union |
| **EP** | European Parliament |
| **EC** | European Commission |
| **Equinet** | European Network of Equality Bodies |

| | |
|---|---|
| **FRIA** | Fundamental Rights Impact Assessment |
| **GDPR** | Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) |
| **GenAI** | Generative Artificial Intelligence |
| **GPAI** | General-Purpose Artificial Intelligence |
| **GREVIO** | Group of Experts on Action against Violence against Women and Domestic Violence |
| **HUDERIA** | Human Rights, Democracy, and the Rule of Law Impact Assessment |
| **ICT** | Information and Communication Technology |
| **ILO** | International Labour Organisation |
| **IoT** | Internet of Things |
| **LLMs** | Large Language Models |
| **ML** | Machine Learning |
| **MSA** | Market Surveillance Authority |
| **NLP** | Natural Language Processing |
| **OECD** | Organisation for Economic Co-operation and Development |
| **DPA** | Data Protection Authority |
| **RAD** | Representative Actions Directive |
| **STEM** | Science, Technology, Engineering, and Mathematics |
| **TEU** | Treaty on European Union |
| **TFEU** | Treaty on the Functioning of the European Union |
| **TFGBV** | Technology-Facilitated Gender-Based Violence |
| **UN** | United Nations |
| **VAW Directive** | Directive (EU) 2024/1385 on combating violence against women and domestic violence |

| | |
|---|---|
| **VLOP** | Very Large Online Platform |
| **VLOSE** | Very Large Online Search Engine |

# LIST OF BOXES

Box 1: Concretisation of risk classification in the AI Act 53
Box 2: Prohibition on AI-generated non-consensual intimate imagery or the so-called nudifier ban 58
Box 3: Fictious example used throughout the chapter 63

# LIST OF TABLES

Table 1: Glossary of key AI-related terms 17
Table 2: Scope of the EU Equality directives 48
Table 3: Overview of bias mitigation measures 64
Table 4: Sample comparison of group fairness metrics 69

# EXECUTIVE SUMMARY

## Context and aim

Artificial intelligence (AI) is often promoted as a catalyst for progress, and in some respects it can support gender equality, for instance by making inequalities in media representation measurable. Yet as AI spreads across all domains of our lives, it can also reproduce and amplify existing gender inequalities and enable new forms of harm, including technology-facilitated gender-based violence (TFGBV). This study analyses, at EU level, the nature, drivers, manifestations, and consequences of gender bias in AI, and recommends action for EU policymakers.

## Scope and method

The study covers AI systems as defined in the AI Act, including general-purpose AI systems, as well as automated decision-making systems that fall outside the Act but reproduce bias in similar ways, in both the public and private sectors. It focuses on the EU and primarily on adult women, reflecting the available evidence. It combines interdisciplinary desk research with a targeted stakeholder consultation and expert review, and covers developments until September 2026. Certain limitations apply, for example, the consultation is not representative, and evidence is rarely disaggregated beyond binary gender.

## Foundational concepts

Bias is a deviation from a standard, so assessing it is a normative choice. The study applies the standard of substantive gender equality, read through an intersectional lens. Gender bias in AI is a systematic deviation from that standard whereby a system, through its data, design or use, disadvantages people on the basis of gender, including gender identity and expression, alone or combined with other characteristics. It need not be intentional or process gender directly. It has three sources: pre-existing bias, rooted in social structures and in the assumptions of those who design and deploy technology; technical bias, arising from choices in problem formulation, data, learning, and evaluation; and emergent bias, which develops after deployment. Beyond gender bias, the study addresses further gendered harms, including TFGBV, an umbrella term for the forms of gender-based violence that technology facilitates, amplifies or newly creates. Like gender bias, TFGBV amounts to gender discrimination, and the two reinforce each other. The study examines TFGBV as a specific, related manifestation of gendered harm.

## Manifestation and impact

The study examines gender bias and gendered harms in these domains:

- *Employment*: bias appears at every stage of recruitment, from job advertising to CV screening and video interviews; algorithmic scheduling and monitoring penalise those with caring responsibilities, and profiling tools have treated care as a characteristic of mothers.
- *Public and private services*: welfare fraud risk-scoring in several Member States disproportionately targeted women, particularly single mothers, women with a migrant background, and low earners; removing gender from credit-scoring models increased discrimination against women.
- *Healthcare*: diagnostic tools are typically developed on a population treated as a universal norm and then applied to others, so accuracy drops for under-represented groups; large language models advised self-management at home more often for women, and AI care summaries downplayed women's needs.

- *Social media*: recommender systems amplify narrow beauty ideals, content moderation removes or downranks women's bodies and sexual and reproductive health information while approving sexually explicit advertising aimed at men, and advertising algorithms reproduce occupational stereotypes.
- TFGBV: takes evolving forms, from image-based sexual abuse, sexual digital forgeries, and sextortion, to cyberharassment, cyberstalking, hate speech, abuse in the metaverse, as well as simulated through chatbots and sex robots. It is fundamentally gendered, intersects with other inequalities, spans online and offline space, causes psychological and economic harm and the silencing of women in public life.

In this context, evidence disaggregating beyond gender, or originating from the EU and its Member States, remains scarce, so that the women most affected are the least visible in the evidence on which policy relies. Women are also under-represented as developers and users of AI.

## Legal, policy and technical responses

Responses range from the EU legal frameworks to prevention, and redress as follows:

- *Legal framework*: EU anti-discrimination law protects unevenly across grounds and domains, and its categories are strained by intersectional and proxy discrimination. The AI Act is highly relevant, but many of its safeguards apply only where an AI system qualifies as high-risk or falls within another regulated category. Consequently, not all instances of AI-related discrimination or gender bias will be captured by its requirements. The GDPR, the DSA and Directive (EU) 2024/1385 add safeguards on personal data, online platforms and certain forms of TFGBV. None systematically integrates a gender-responsive approach, leaving protection dependent on interpretation and enforcement. Simplification initiatives, such as the withdrawal of the AI Liability Directive and the Digital Omnibus package, raise questions about how equality objectives will be balanced against competitiveness. At the same time, the AI Act's maximum harmonisation through product safety legislation sits uneasily with broader national fundamental rights protection.
- *Prevention*: AI systems are socio-technical systems, so their outcomes depend on technical design, organisational practices, deployment contexts, and broader social structures. Accordingly, technical mitigation is limited: proxies persist, fairness metrics embody different conceptions of equality, and correcting one bias can create another. The AI Act does not require bias to be eliminated: providers of high-risk systems must mitigate foreseeable risks and then judge themselves, through conformity assessment, whether the residual risk is acceptable. As harms often arise in deployment, certain deployers, notably public authorities, must assess fundamental rights impacts before use, which may support a decision not to deploy. These assessments are, however, limited in scope and do not require the participation of affected groups, and gender impact assessments, long used in EU gender mainstreaming, are absent from the digital acquis. The meaningful involvement of women in the design, governance and assessment of AI remains essential.
- *Redress*: algorithmic discrimination is hard to detect, attribute and prove, so scrutiny often relies on external actors, and its intersectional, systemic and collective nature exceeds individual remedies. Responsibility and liability are fragmented across complex, cross-border value chains. Despite progress, such as through enhanced transparency rights for affected individuals and the involvement of specialised actors in oversight and enforcement, including fundamental rights authorities under the AI Act and trusted flaggers under the DSA, gaps remain: the AI Act's complaint mechanism lacks the guarantees of the GDPR, collective redress

is largely confined to consumer law, and no EU instrument addresses the evidentiary barriers, leaving the burden-of-proof rules of anti-discrimination law central.

## Recommendations

The study recommends actions on four priorities: (1) strengthening the gender equality framework, including through the adoption of the Horizontal Equal Treatment Directive, so that it can address intersectional, systemic and collective algorithmic harm; (2) embedding a gender-responsive and intersectional approach in the development and application of the digital acquis; (3) ensuring well-resourced enforcement and oversight, including by equality bodies; and (4) improving evidence collection and the participation of affected groups, with dedicated funding for their representatives. Collectively, these measures seek to close existing protection gaps, strengthen accountability, and ensure that gender equality is treated as a core objective of AI governance rather than a peripheral consideration.

# 1. INTRODUCTION

Artificial intelligence (AI) is often promoted as a catalyst for progress and, in some respects, it has contributed to the pursuit of gender equality. AI can, for instance, make gender inequalities visible at a scale that manual analysis cannot achieve. A machine learning (ML) tool developed for the Geena Davis Institute on Gender in Media analysed the 200 highest-grossing films of 2014 and 2015 and found that male characters received roughly twice as much screen time as female characters (Geena Davis Institute on Gender in Media, 2016). Similarly, automated hate speech detection can help online platforms identify and remove misogynistic content at scale.[1] Yet, AI can also reproduce and amplify existing gender inequalities, and it is this risk that the study addresses.

## 1.1. Research questions

The general objective of the study is to analyse, at the European Union (EU) level, the nature, drivers, manifestations and consequences of gender bias in AI. The study approaches gender bias in AI from the angle of substantive equality and intersectionality. Rather than treating bias as a technical flaw to be corrected, it regards it as a socio-technical phenomenon through which existing gender inequalities are reproduced, amplified, or newly created, and through which technology can also be used to commit violence against women. Accordingly, the study addresses the following main research question: To what extent are the EU's legal, policy and governance frameworks equipped to prevent, mitigate, and redress gender bias in AI, and what action should EU policymakers take to close the gaps?

To answer it, the study examines six sub-questions.

1. How should gender bias in AI be understood, and where does it originate? (Chapter 2)
2. How does gender bias in AI affect women, and which groups of women in particular, across key domains of life? (Chapter 3)
3. How does technology, including AI, facilitate gender-based violence, and how does this relate to gender bias? (Chapter 4)
4. How do EU non-discrimination laws and the digital acquis address gender bias in AI, and where do gaps and tensions arise? (Chapter 5)
5. Which technical, organisational, and governance measures are available to prevent and mitigate gender bias, and what are their limits? (Chapter 6)
6. How effective are existing redress, oversight and accountability mechanisms in addressing algorithmic discrimination? (Chapter 7)

The study proceeds from three working hypotheses. First, gender bias in AI is not primarily a technical malfunction but a reflection of structural inequality, so that technical measures alone cannot remove it. Second, gender bias and technology-facilitated gender-based violence (TFGBV) are distinct but mutually reinforcing forms of gender discrimination, shaped by the intersection of gender with other characteristics, which an effective response must address together. Third, anti-discrimination laws and the EU digital acquis were developed largely independently of one another and pursue distinct regulatory objectives. While each framework provides important safeguards, both face limitations in addressing the systemic, societal, intersectional and often difficult-to-detect nature of algorithmic harm.

[1] However, AI use for this purpose remains contested and subject to limitations, such as difficulties in interpreting context and uneven accuracy across languages (Zeinert et al., 2021).

## 1.2. Methodology

The study adopts a qualitative, multi-method research design that combines desk research with a targeted stakeholder consultation. The desk research covered academic literature, grey literature from EU, international institutions and civil society, investigative reporting, and legislation and case-law. The consultation comprised a questionnaire, disseminated among others through the European Network of Equality Bodies (Equinet) to European equality bodies. A detailed account of the methodology, its limitations and the authors' positionality is provided in Annexes 1, 2 and 3.

## 1.3. Structure of the study

The study is composed of seven substantive chapters. Chapter 2 sets out its conceptual foundations. It defines gender bias against the standard of substantive equality, develops an interdisciplinary understanding of gender and intersectionality, and maps the principal sources of gender bias in AI systems, from biased datasets to insufficient diversity in design processes.

Chapter 3 and 4 examine harms. Chapter 3 examines how gender bias manifests across key domains of life, including employment, social media, healthcare, and credit scoring. Drawing on case studies and lived experiences, it demonstrates that gender bias in AI disproportionately affects women, and some groups of women in particular, generating economic, health-related, psychological, and representational harms. Chapter 4 turns to technology-facilitated gender-based violence (TFGBV), itself recognised as a form of discrimination against women, and analyses how different technologies, including AI, facilitate, amplify, and create new manifestations of gendered harms.

Chapters 5 to 7 shift the focus from harms to responses, and evaluate the extent to which existing legal, technical, and governance responses are capable of addressing the systemic, intersectional, and often difficult-to-detect nature of algorithmic harm. Chapter 5 maps the EU legal and policy framework, focusing on the interaction between the digital acquis and anti-discrimination law. The chapter assesses both recent legislative developments and the challenges arising from their implementation and interaction. Chapter 6 examines the socio-technical, organisational, and governance measures available to prevent and mitigate gender bias throughout the AI lifecycle. Chapter 7 analyses the effectiveness of the redress, oversight, and accountability mechanisms.

On this basis, Chapter 8 answers the main research question, synthesising the findings into a set of recommendations for EU policymakers and identifying priority actions for the European Parliament (EP) and the FEMM Committee.

## 1.4. Scope

Three deliberate choices define the breadth scope of the study. First, it is not limited to AI systems as defined in the Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on AI (AI Act) but also covers ADM systems. While some ADM systems lack the level of adaptiveness and autonomy required to qualify as AI systems, they are nonetheless important to examine, given their discriminatory potential (Xenidis, 2025). Similarly, the study examines TFGBV in its broader form, including manifestations that do not involve AI, since AI is increasingly one of the technologies through which such violence is committed, and the line between AI and other technologies is not stable.

Second, the study addresses AI and ADM systems across both private and public sectors. Given the objective of examining the nature, drivers, manifestations, and consequences of gender bias in AI at

the EU level, limiting the analysis to either the public or private sector would unduly narrow its scope. Discriminatory risks may arise during the design and development of AI systems, but only become apparent once those systems are deployed in practice. The study therefore reflects the interdependence between private-sector development and public- or private-sector deployment (Smuha, 2025).

Third, while gender bias is defined inclusively when it comes to different gender identities and expressions, the study analysis centres primarily on women, because the available evidence rarely disaggregates findings beyond binary gender. This evidence gap is itself a finding of the study. Finally, the study focuses on the risks that AI poses to gender equality rather than on its potential to advance it, which is addressed only briefly.

# 2. AI AND GENDER BIAS: CONCEPTS, DEFINITIONS AND SOURCES

KEY FINDINGS

- There is no single, universally accepted definition of AI. Most official definitions draw on a few sources, notably the Organisation for Economic Co-operation and Development (OECD), which has encouraged consistency but also carried ambiguities from one framework to the next.
- The EU definition has moved from a broad, strategic description to a legal one built around inference. Under the AI Act, an AI system is defined by its capacity to infer how to generate outputs, which distinguishes it from rule-based software, and whether a system qualifies must be assessed case by case.
- Bias has no settled meaning in EU law, policy or scholarship. The AI Act uses the term without defining it, and also uses the term in several senses, ranging from bias in datasets to automation bias in human oversight.
- Whether an AI system is biased depends on the standard it is measured against, and choosing that standard is a normative decision. If a system is judged only by how accurately it reflects the world as it is, existing gender inequalities are treated as accurate rather than biased.
- Gender bias can enter an AI system at every stage of its lifecycle. It can come from social inequalities recorded in data, from choices made in design and development, or from changes after deployment, and these sources are interconnected.
- Removing gender from AI systems does not make them gender-neutral, but including gender is no simple remedy either. AI systems can reconstruct gender from correlated variables, while collecting gender data to measure bias raises privacy risks and forces choices about categories.

Concern about gender bias in AI has grown as AI systems have spread across many areas of life, where they can reproduce or amplify existing gender inequalities. Addressing that concern, however, depends on two concepts that remain contested: AI and gender bias. Accordingly, this chapter clarifies how AI is defined in the literature and at the EU level (Sections 2.1 and 2.2), sets out the normative standard against which this study assesses gender bias in AI (Section 2.3), and maps the sources of such bias (Section 2.4).

## 2.1. The concept of artificial intelligence

There is no universally accepted definition of AI. As Narayanan and Kapoor argue, AI 'is an umbrella term for a set of loosely related technologies' (2024, p. 1), much as 'vehicle' would be if it were the only word available for bicycles, buses, cars, and spacecraft alike.

Nonetheless, scholars and policymakers have sought to define AI systems, largely because regulation cannot function without clear terms. Through a systematic review of 896 academic papers and a manual review of more than 80 regulatory, standards, and policy documents, Sun et al. (2026) find that most official definitions stem from a small number of sources, above all Russell and Norvig's textbook *Artificial Intelligence: A Modern Approach*, and the work of the OECD on AI. This shared lineage has

encouraged a degree of international consistency, but it has also carried unresolved ambiguities from one framework to the next, such as the distinction between an 'AI system' and an 'AI model' (Sun et al., 2026), discussed below.

Before turning to how the EU understands and defines AI, the table below clarifies some related concepts that are often associated with AI and that will be used throughout this study:

Table 1: Glossary of key AI-related terms

| Term | Definition |
|---|---|
| Algorithm | A finite sequence of well-defined instructions that a computer follows to solve a problem or complete a task. Algorithms underpin all software, not only AI (Steckelmacher and Bogaerts, 2025). |
| ADM system | Systems that take decisions without human involvement (Article 29 Working Party, 2018). ADM does not necessarily require AI as it could follow fixed rules. |
| Deep learning | A subset of ML that uses artificial neural networks with many layers to learn increasingly abstract features from data. Each layer transforms its input and passes the result to the next, allowing the system to handle complex tasks, such as recognising faces in images or understanding speech (Burgard, 2022). |
| Generative AI (GenAI) | AI systems that produce new content, such as text, images, audio, video, or computer code, in response to a prompt. Rather than classifying or predicting from existing data, they generate new outputs based on patterns learned during training (Fernández-Llorca et al., 2024; Lorenz et al., 2023; Narayanan and Kapoor, 2024). Classic examples include large language models (LLMs) such as ChatGPT and Claude, and image generators such as DALL·E and Midjourney. |
| Machine Learning (ML) | A branch of AI in which systems improve their performance on a task by learning from data rather than following explicitly programmed rules. A model is trained on examples and can then make predictions or decisions when presented with new data (Burgard, 2022; Steckelmacher and Bogaerts, 2025). |
| Natural language processing (NLP) | A branch of AI concerned with enabling computers to process, interpret, and generate human language, whether written or spoken (Steckelmacher and Bogaerts, 2025). Applications include machine translation, speech recognition, chatbots, and text summarisation. |

Source: Authors' compilation, on the basis of cited literature.

## 2.2. The definition of AI at the EU level

The European Commission (EC) first engaged with the question of how to define AI in 2018, with its Communication on *'Artificial Intelligence for Europe'*. The definition it offered was deliberately broad and served a strategic rather than legislative purpose (EC, 2018). In 2020, the EC's *White Paper on Artificial Intelligence* marked a shift towards regulation, and warned that any future legal definition would need 'to be sufficiently flexible to accommodate technical progress while being precise enough to provide the necessary legal certainty' (EC, 2020b, p. 16).

The tension between flexibility and precision became central when the EC proposed the AI Act in April 2021. The proposal defined an AI system by reference to a list of techniques set out in Annex I, an

approach criticised for being so broad that it could create legal uncertainty (Ruschemeier, 2023; Veale and Zuiderveen Borgesius, 2021). During the legislative process, the co-legislators abandoned the list-based approach and converged on the definition revised by the OECD in November 2023 (Fernández-Llorca et al., 2024; Zanfir-Fortuna, 2026).

The final text of the AI Act, adopted as Regulation (EU) 2024/1689, defines an AI system as: 'a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and 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' (Article 3(1)). The defining feature of this formulation is the capacity to 'infer'. Recital 12 of the AI Act explains that this capacity distinguishes AI systems from simpler traditional software and from programming approaches that are based solely on rules defined by human beings to execute operations automatically.

As the definition left considerable room for interpretation, the EC published guidelines on the definition of an AI system in February 2025. The guidelines break the definition down into seven elements, give examples of systems that fall outside it, such as those limited to basic data processing, and emphasise that the assessment must be made on a case-by-case basis (EC, 2025a).

The AI Act contains two further definitions relevant to this study, introduced after the public release of ChatGPT in November 2022 drew attention to general-purpose and generative systems (Fernández-Llorca et al., 2024). Under Article 3(63), a general-purpose AI model is one that 'displays significant generality' and can 'competently perform a wide range of distinct tasks', and that can be integrated into a variety of downstream systems or applications. Under Article 3(66), a general-purpose AI system is a system built on this model that is itself capable of serving a variety of purposes.

Accordingly, the model is a component of the system rather than a synonym for it (Zanfir-Fortuna, 2026). The distinction carries legal weight, since the AI Act imposes obligations not only on those who place AI systems on the market but also, in Chapter V, on the providers of general-purpose AI models. As Sun et al. (2026) observe, this allocation of responsibilities across the value chain only works if the boundary between model and system can be drawn with reasonable clarity, which is precisely where the existing definitions remain ambiguous.

In light of the above, this study adopts the definition of an AI system in Article 3(1) AI Act and that of a general-purpose AI system in Article 3(66) AI Act. As explained in the Introduction, it also covers ADM systems which, where they rely solely on human-defined rules, fall outside the AI Act (EC, 2025a), but can reproduce and amplify gender bias in similar ways.

## 2.3. The concept of gender bias in AI

'Bias' is one of the terms most frequently invoked in debates on AI and one of the least often defined.

The closest the EU has come to defining bias is the glossary of the non-binding *Ethics Guidelines for Trustworthy AI* (2019), drawn up by the High-Level Expert Group on Artificial Intelligence (AI HLEG). The glossary describes 'bias' as 'an inclination or prejudice for or against a person, object, or position' (AI HLEG, 2019b, p. 36), which may enter through the data, the design of a system, or the way it is used, and which is not necessarily undesirable (AI HLEG, 2019b).

The AI Act refers to bias in several provisions without defining it, and the term does not carry the same meaning throughout. Article 10(2)(f)–(g) concerns biases in datasets that are likely to affect health, safety, or fundamental rights, or to lead to discrimination prohibited under EU law. Article 15(4) addresses biased outputs of systems that continue to learn after deployment and may feed back into

their future operation (Meding, 2025). Article 14(4)(b), by contrast, uses the term in a cognitive sense. It requires that those overseeing high-risk systems be able to remain aware of 'automation bias', which is the tendency to rely, or over-rely, on the output an AI system produces (Laux and Ruschemeier, 2025). As a result, the AI Act moves between understandings of bias located in data, in the system, and in human judgement, without signalling the shift. Nonetheless, more precise, technical definitions are expected to come from the harmonised standards being developed by the European Committee for Standardization (CEN) and the European Committee for Electrotechnical Standardization (CENELEC), which are expected to draw on ISO/IEC TR 24027:2021 to define bias in AI systems (EC, 2023; Soler Garrido et al., 2024).

This definitional gap runs through academic literature as well (Puttick et al., 2025). The different disciplinary meanings of the term can be connected by understanding bias, in its most general form, as a systematic deviation from a standard, which may be statistical, moral, legal, or social (Danks and London, 2017; T. Kelly, 2022). In this abstract sense, the concept is neutral: whether a deviation is a problem depends on the standard chosen.

A model that associates nursing with women can illustrate the point. Measured against the current composition of the profession, the association may be statistically accurate. Measured against a norm under which occupation should not be linked to gender, the same association is biased. The model is the same in both cases. What differs is the reference point: the world as it is, or the world as it ought to be (Mitchell et al., 2021). In AI, however, the choice between these reference points is seldom neutral because the data from which systems learn and the standards against which they are assessed are themselves shaped by the social structures and hierarchies they sustain. Accordingly, the choice of a standard is a normative act.

In this study, the normative standard against which bias is assessed is substantive gender equality. The concept is best understood in contrast with formal equality, which requires that like cases be treated alike: a rule or system is equal if it applies the same criteria to everyone, regardless of gender. Instead, substantive equality looks at outcomes and at the structures that produce them, recognising that identical treatment can entrench disadvantage where people start from unequal positions. According to Sandra Fredman (2016), it has four dimensions: redressing disadvantage; addressing stigma, stereotyping, prejudice and violence; enhancing voice and participation; and accommodating difference through structural change. This understanding is reflected in EU law, as mentioned in Chapter 5. In the context of gender bias in AI, the distinction is decisive. As Section 2.4.2 will show, the removal of gender from a system's data does not make it gender neutral, while fairness measures that preserve existing inequalities sit uneasily with the substantive aims of non-discrimination law (Wachter et al., 2021). Accordingly, the assessment of AI systems against substantive equality asks not only whether they treat people alike, but whether they reproduce, amplify, or help redress gendered disadvantage.

The application of this normative standard requires clarity on two further concepts, namely gender and intersectionality. Gender is understood here as a psychological, cultural, and social construction rather than a biological given. It is shaped through socialisation and is closely tied to personal identity, that is, to how individuals experience and understand themselves (Butler, 2007). Gender does not, however, operate in isolation. The theory of intersectionality, introduced by Kimberlé Crenshaw (1989), captures how gender interacts with other social categories. Drawing on US anti-discrimination case-law, Crenshaw showed that the separate evaluation of sex and race takes the most privileged members of each group, i.e. White women and Black men, as the norm, rendering invisible those at the intersection, such as Black women (Crenshaw, 1989).

The theory of intersectionality has since moved well beyond its legal origins (Hill Collins and Bilge, 2020). It is particularly relevant to the study of bias in AI, because computational methods depend on categories. It raises questions about which categories are used, how they are combined, and whether analysing them one at a time conceals patterns that appear only at their intersection. These patterns become visible when an AI system's performance is broken down by combined characteristics (Weerts et al., 2023). An audit of commercial facial analysis systems, for example, found that error rates were lowest for lighter-skinned men and highest for darker-skinned women. Examining gender or skin type separately would have understated this gap (Buolamwini and Gebru, 2018). Chapter 3 examines these cases and their impact on women and society in more detail.

Against this background, this study considers gender bias in AI as a systematic deviation from the standard of substantive equality irrespective of gender, whereby an AI system, through its data, design, or use, produces outputs or effects that disadvantage individuals or groups on the basis of their gender, including their gender identity and expression, alone or in combination with other characteristics and social inequalities. Gender bias does not need to be intentional, nor does it require that gender be processed directly. It covers both the reproduction of existing gender inequalities and their amplification, and it may affect people of any gender, including women and men, whether cis or trans, as well as intersex, non-binary, gender non-conforming, gender-fluid, and queer individuals. All of these groups face persistent structural inequalities, which AI systems can reproduce and amplify. For trans, intersex, non-binary, and gender non-conforming persons, these inequalities are compounded where AI systems treat gender as binary and fixed, or infer it automatically from names, faces, or voices. However, the available evidence focuses overwhelmingly on women and rarely disaggregates findings beyond binary gender. As a result, the analysis that follows primarily focuses on cis and trans women, while addressing the experiences of other groups wherever evidence allows, and treats this evidence gap as a finding in its own right.

Gender bias is not the only harm that the standard of substantive equality brings to light. It also captures TFGBV, examined in Chapter 4. Violence is explicitly named in the second of Fredman's dimensions, but TFGBV, like gender bias, undermines the other dimensions too. Both cause disadvantage, whether through biased decisions on employment, services and healthcare or through the psychological and economic costs of abuse. Both also constrain women's voice and participation, through reduced visibility online or withdrawal from public life (Chapters 3 and 4). Moreover, the two reinforce each other. The stereotypes that AI systems reproduce normalise the objectification on which much TFGBV relies, and the same class of generative systems is used to create sexual digital forgeries (Section 4.2.2). Conversely, gender bias can weaken systems designed to respond to violence (Section 3.5), and abusive content can feed the data from which systems learn (Section 2.4.3). Therefore, gender bias and TFGBV are mutually reinforcing obstacles to substantive equality.

## 2.4. The sources of gender bias in AI

One of the most influential frameworks for understanding how bias arises in AI systems is that of Batya Friedman and Helen Nissenbaum (1996), who distinguish three types of bias according to where it originates: pre-existing, technical, and emergent. This study adopts their framework because it identifies the origin of bias, and thereby the actors and moments at which it must be addressed. The categories are analytical rather than mutually exclusive. They also operate on a different axis from the stages of development identified by Suresh and Guttag (2021), which describe where bias enters an AI system. The two lenses are complementary. Pre-existing bias, which arises from society, usually enters an AI system through technical stages such as data collection or labelling, and technical choices determine whether it is reproduced, amplified, or corrected. Conversely, technical choices are never

purely technical, since they are made by people whose assumptions are shaped by pre-existing bias. Accordingly, this section uses the three categories to identify where gender bias comes from, and the stages of development to trace how it enters the system. In practice, several sources of bias may contribute to the same outcome.

### 2.4.1. Pre-existing bias

Pre-existing bias originates outside the technical system. It stems from social structures, norms, and institutions, both public and private, as well as from the assumptions of those who design, develop, and deploy technology. It can be introduced deliberately or without awareness (Friedman and Nissenbaum, 1996).

At the societal level, pre-existing bias often takes the form of historical bias. This arises when patterns of past inequality are recorded in the data, so that even perfectly collected and measured data will reproduce them (Suresh and Guttag, 2021). An experimental recruitment tool developed by Amazon in the 2010s is a classic example. Trained on ten years of CVs submitted to a male-dominated technical workforce, it learned to favour male candidates and reportedly downgraded CVs containing the word 'women's', as in 'women's chess club captain' (Dastin, 2018). Language technologies show similar patterns. Word embeddings, which represent words as numerical vectors and place words used in similar contexts close together, are learned from large collections of text and, consequently, they take on whatever associations that text contains. Embeddings trained on web content, for example, tie men more closely with careers and women with family, closely mirroring associations measured in people (Caliskan et al., 2017). In each of these cases, the systems reproduce patterns in their data, demonstrating that accurate data and neutral data are not the same thing.

At the individual level, pre-existing bias stems from the assumptions of those who design, develop, and deploy an AI system. These assumptions shape design choices even where no one intends to favour a particular group. The female names, voices, and deferential personalities given to voice assistants such as Siri, Cortana, and Alexa are often cited as a design choice reflecting gendered assumptions about service roles (Adams and Loideáin, 2019). Furthermore, research on earlier information and communication technologies (ICTs) suggests one reason why these assumptions go unexamined: developers tend to take themselves as representative of future users (Oudshoorn et al., 2004). This means that the composition of design teams shapes whose perspectives are built into an AI system. Accordingly, the current under-representation of women in the AI workforce affects which assumptions are noticed and which go unquestioned (European Institute for Gender Equality (EIGE), 2021; West et al., 2019).

### 2.4.2. Technical bias

Technical bias arises from choices and constraints in the design and development process itself (Friedman and Nissenbaum, 1996).

The first point at which technical bias can arise is problem formulation. Turning a social question into a prediction task requires deciding what the AI system should predict and which measurable variable will stand in for an abstract concept such as a 'good employee', a 'reliable borrower', or a 'patient in need of care'. These choices are value-laden (Passi and Barocas, 2019). The ML system used by Sweden's Social Insurance Agency illustrates the point. It was applied to the temporary parental allowance, which compensates parents who stay home to care for sick children, and it treated benefit administration as a matter of predicting fraud rather than, for instance, identifying errors. Applicants with the highest scores were referred directly to the agency's fraud control unit, which treated cases as involving suspected criminal intent. Because women take the majority of this leave and are more likely to combine

it with part-time or irregular work, the irregularities the system treated as signs of fraud were more likely to arise in their claims. As a result, the way the problem was formulated directed suspicion towards the caregiving patterns typical of women. A journalistic investigation found that the system disproportionately flagged women, as well as people with a migrant background, low-income earners, and people without a university degree (Granberg et al., 2024).

Even where a problem is carefully formulated, gender bias can arise from the data used to address it. Representation bias occurs when the data under-represents or over-represents some groups relative to the population that the AI system will serve (Suresh and Guttag, 2021). The facial analysis datasets examined by Buolamwini and Gebru (2018) contained comparatively few images of darker-skinned women. Similar gaps recur across domains, from automatic captioning, which is less accurate for women's voices (Tatman, 2017), to medical imaging (see Section 3.3.1). Carolina Criado-Perez (2020) calls the underlying pattern a 'gender data gap': the long-standing tendency to collect data on men and to treat the male as the default, which AI systems inherit whenever they learn from these records. Intersectional subgroups are particularly exposed, because a dataset can appear balanced by gender and by other personal characteristics taken separately, while containing very few people who belong to a particular combination.

Nonetheless, numbers alone do not guarantee accurate representation. A dataset may include women in adequate proportions and still misrepresent them because measurement bias stems from the way features and labels are defined, measured, and recorded (Suresh and Guttag, 2021). Features often capture more than they appear to. The profiling system used by the Austrian Public Employment Service exemplifies this. Among the variables used to predict job seekers' chances of re-employment, a variable for care obligations lowered predicted chances, but this applied only to women. Its developers defended this choice as reflecting the unequal distribution of care work (Achterhold et al., 2025). By recording care as a characteristic of women alone, the system encoded a gendered division of labour into its measurements, at the intersection of gender and caring responsibilities. Labels raise a similar problem. Outcomes such as performance ratings or past hiring decisions reflect the judgements of those who assigned them, meaning that an AI system trained to reproduce them inherits those judgements (Barocas and Selbst, 2016).

Finally, gender is often recorded as a binary field or inferred automatically from appearance, building a particular and contestable model of gender into the data (Fosch-Villaronga et al., 2025; Keyes, 2018). Where measurement bias concerns how groups are categorised, aggregation bias arises when meaningful differences between them are ignored. It occurs when a single model is applied to groups whose underlying patterns differ: that model fits the majority and misfits others (Suresh and Guttag, 2021). The diagnosis of heart attack shows how this works. It relies on blood levels of the protein troponin, compared against a single threshold. Because women typically have lower troponin levels than men, a uniform threshold set largely on male patients can miss heart attacks in women. In one large study, the application of sex-specific thresholds roughly doubled the number of women diagnosed (Shah et al., 2015). The same risk arises when ML models are trained on pooled patient data with a single decision threshold, which is why recent diagnostic algorithms incorporate sex explicitly (Doudesis et al., 2023).

Gender bias can also emerge from the learning process. Learning bias arises when choices made in training a model, including the objective it is optimised for, amplify disparities rather than simply reproduce them (Suresh and Guttag, 2021). Text-to-image generators have, for instance, been found to exaggerate occupational and social stereotypes (Bianchi et al., 2023). In a recent study of three popular systems, namely, Midjourney, DALL·E, and Stable Diffusion, adding the word 'competent' to a prompt made the depiction of a man nine times more likely, and one of them produced no images of

women at all (Weinmann et al., 2026). Objectives that appear neutral can have the same effect: ad delivery algorithms that minimise cost have shown job advertisements to women and men in gender-stereotyped ways, even when advertisers did not target by gender (see Section 3.1.2). Once a model is trained, gender bias can persist undetected if it is tested in the wrong way. Evaluation bias arises when the benchmarks and metrics used to test a system do not reflect the population or conditions it will face (Suresh and Guttag, 2021). For example, commercial machine translation systems that scored well on conventional quality metrics were found to be systematically prone to gender errors when translating English sentences into eight languages, including French, German, Italian and Spanish, particularly where a person's role contradicted stereotypes (Stanovsky et al., 2019).

A common assumption, underlying many responses to these sources of bias, is that gender bias arises because systems process gender directly, suggesting that removing gender and other protected characteristics from the data will prevent it. This approach is sometimes called 'fairness through unawareness' (Kusner et al., 2017). Research suggests that it offers little protection. Where sensitive attributes are excluded, AI systems can reconstruct them from correlated variables such as first names, career breaks, educational background, purchasing patterns or hobbies (Wachter et al., 2021), which Barocas and Selbst (2016) call 'redundant encodings'. The experimental Amazon recruitment tool discussed above shows this: it had no gender field, yet gendered words in CVs served as a stand-in. Prince and Schwarcz (2020) argue that this proxy discrimination is a structural feature of AI, because systems trained to find predictive correlations will tend to find proxies whenever a protected characteristic is itself predictive of the outcome.

Conversely, processing sensitive data may be necessary to identify and correct bias, since disparities between groups cannot be measured without knowing who belongs to which group (Žliobaitė and Custers, 2016). This data collection, however, raises difficulties of its own. It exposes individuals to privacy risks and depends on their willingness to disclose personal information. It also requires choosing categories, which returns to the question of how gender is defined, since binary options cannot capture non-binary identities. Intersectional analysis amplifies these problems because it requires data on several characteristics at once, increasing the privacy risks and the difficulty of obtaining enough data for each subgroup. Practitioners report these obstacles as significant constraints on bias assessment (Andrus et al., 2020). In conclusion, excluding gender from AI systems does not make them gender-neutral, but including it is no simple remedy either.

### 2.4.3. Emergent bias

Unlike the first two types, emergent bias develops after design, when an AI system is used in new contexts or when social norms, values, and expectations change (Friedman and Nissenbaum, 1996). AI systems built for one setting are sometimes transferred to another whose social conditions differ, a risk described as the 'portability trap' (Selbst et al., 2019). Changing norms can have a similar effect. Many AI systems encode gender as a fixed, binary attribute, often inferred from names, faces, or voices (Keyes, 2018). When these AI systems were designed, this may have reflected dominant norms, including legal and administrative categories. Those categories have since changed in several respects, including through reforms of legal gender recognition across the EU, some of which now recognise options beyond 'male' and 'female' (EC, 2020a). Accordingly, a system built around binary, immutable gender categories becomes increasingly out of step with the legal and social reality in which it operates, without any change to the system itself. Gender bias can also emerge from the way an AI system continues to learn once deployed. AI systems that learn from the outcomes of their own decisions can generate feedback loops, in which an initial pattern is reinforced over time (Ensign et al., 2017), as in the engagement-based recommendation of sexualised images of women on Instagram (see Section 3.4.1). Users can also introduce bias directly. In 2016, Microsoft's chatbot Tay began producing

misogynistic and racist content within a day of its release, having learned from users who deliberately fed it such material (Neff and Nagy, 2016), and similar patterns have since been observed in more recent chatbots (McGlynn et al., 2026).

### 2.4.4. Concluding considerations

The value of this categorisation lies less in sorting individual biases into boxes and more in what each category implies for identifying and addressing gender bias in AI, as Chapter 6 will further elaborate. In brief, pre-existing bias cannot be detected by checking whether a system accurately reflects its data, since accurate data may themselves encode inequality, as the nursing example in Section 2.3 shows. Its identification therefore depends on a normative choice, which is built into fairness metrics themselves. Pre-existing bias also cannot be removed by technical means alone and calls for measures that address the composition of people and processes that feed into the system. By contrast, technical bias arises from identifiable design and development choices, and forms the main target of the AI Act's risk management and data governance requirements, and of technical mitigation measures. Finally, emergent bias cannot be fully anticipated before deployment and requires continuous monitoring and assessment by those who deploy the system, effective oversight, and redress once harm occurs. Taken together, these categories show that addressing gender bias across the AI lifecycle requires a combination of technical, organisational, governance and legal measures, discussed in Chapter 6.

# 3. HOW GENDER BIAS IN AI AFFECTS WOMEN: EVIDENCE FROM FOUR DOMAINS

KEY FINDINGS

- Gender bias in AI affects women in all four examined domains: employment, services, healthcare, and social media. AI systems tend to reproduce, and sometimes amplify, the inequalities already present in each domain.
- Women are under-represented among those who build and use AI. They make up around 30% of the AI workforce and use GenAI tools less than men, even in the same occupations and firms.
- Gender bias can enter at every stage of employment. It has been found in job ad delivery, CV screening, and video interviews. Algorithmic management can also penalise caring responsibilities, which still fall mainly on women.
- Welfare risk-scoring systems in several Member States have targeted women, particularly single mothers and women with a migrant background. Because public authorities exercise coercive power, the consequences of being wrongly flagged can be severe.
- Removing gender from financial models does not make them neutral. Other variables can act as proxies. In credit scoring, this has reduced women's chances of obtaining credit even though they default less often than men.
- AI in healthcare inherits medicine's treatment of the male body as the norm. Diagnostic models can miss more cases in women, and language models have downplayed women's care needs and advised them to manage conditions at home more often than men.
- Femtech reduces the diversity of women's bodies and lives to a single norm. This excludes many users and reduces the accuracy of predictions.
- Social media algorithms amplify narrow beauty ideals and apply double standards to women's bodies. Content about women's bodies and health is often removed or hidden, while sexually explicit ads aimed at men are approved.
- The women most affected by AI are often the least visible in the evidence. Most studies come from outside the EU, and few disaggregate their findings beyond gender.

This chapter analyses how gender bias in AI systems affects women, paying special attention to how gender intersects with other personal characteristics, such as race, migration background, age, disability, sexual orientation, socio-economic status, and caring responsibilities. Gender bias is far from being new, but AI gives it new dimensions: compared with human or manual decision-making, AI systems can produce discriminatory outcomes at far greater scale and speed, and with less transparency. Although, as explained in Section 2.3, gender bias may affect women, men, and non-binary people, this chapter primarily focuses on adult women. It focuses on four domains: employment and the labour market (Section 3.1), public and private services (Section 3.2), healthcare (Section 3.3), and social media (Section 3.4). These domains were selected because AI is already widely deployed in them, because they shape women's access to income, essential services, health and visibility, and because they have attracted the most research and investigations to date.

Each section follows a similar structure. It first explains how AI is being introduced into the domain and how that domain is already shaped by gender inequalities, before examining specific AI systems and the evidence of their impact on women. Where possible, the analysis draws on cases from EU Member States. Much of the available evidence, however, still comes from the United States and the United Kingdom, and few studies disaggregate their findings beyond gender. This scarcity of intersectional and European data is itself a finding: the women most affected by AI are often the least visible in the evidence on which policy relies.

The analysis is not exhaustive: Section 3.5 closes the chapter with examples from other domains, such as policing and border control, where AI also has gendered effects, while TFGBV is separately examined in Chapter 4.

## 3.1. Employment and the labour market

AI is changing the tasks workers perform (Bujold et al., 2023), and GenAI is accelerating this shift (Aguinis et al., 2024). Across the European Union, employers also increasingly use AI systems to rank job applicants, allocate shifts, monitor productivity, and inform promotions (Paliokaitė et al., 2025). These changes are unfolding in a labour market that is already deeply gendered, meaning that AI risks reproducing, and in some cases amplifying, existing inequalities within society.

### 3.1.1. Exposure to AI and equal access

Women's exposure to AI depends on the type of AI considered. When measuring how far the abilities required in each occupation overlap with the current capabilities of AI, the OECD (OECD, 2024) finds that women and men face similar overall exposure. However, women are under-represented in the occupations most exposed to AI, often high-skilled roles where AI is more likely to augment work, which may limit their access to the opportunities it creates (OECD, 2024). GenAI is a different story. An International Labor Organization (ILO) research brief (ILO, 2026) combining a global index of occupational exposure to GenAI with labour force microdata from 84 countries finds that 29% of female-dominated occupations are exposed to GenAI, compared with 16% of male-dominated ones. The gap widens at the highest level of exposure, which includes 16% of female-dominated occupations but only 3% of male-dominated ones. The ILO attributes this to women's concentration in clerical, administrative, and business support roles involving routine tasks. Women are more exposed than men in 88% of the countries covered, and Europe is among the regions with the highest average exposure for female workers (ILO, 2026).

However, exposure does not equate to job loss. It measures how far an occupation's tasks are likely to be affected, whether through augmentation, restructuring or automation (Fernandez et al., 2026). Evidence from workplaces that have adopted AI points in the same direction. Based on nearly 100 case studies in manufacturing and finance across eight countries, including Austria, France, Germany, and Ireland, the OECD (OECD, 2023) finds that AI has so far led more often to the reorganisation of jobs, shifting them towards tasks where people retain a comparative advantage. Similarly, the ILO expects the wider impact to fall on job quality rather than job numbers, with GenAI potentially intensifying workloads, reducing autonomy, and introducing bias (ILO, 2026).

Importantly, women's exposure to AI is not a technological accident. It reflects entrenched occupational segregation, discriminatory norms, and biases in recruitment and employment that have channelled women into particular roles (ILO, 2026). If adaptation to AI mainly benefits more educated and higher-income workers, it will deepen these existing inequalities (OECD, 2024).

Women are under-represented among both AI developers and AI users at work (OECD, 2024), and made up only about 30% of the global AI workforce in 2022, up just four percentage points since 2016

(ILO, 2026). Stereotypes and the digital skills divide limit women's entry into AI careers, while male-dominated environments, sexual harassment, and limited access to funding push them out later (EIGE, 2021). At the same time, a usage gap is emerging. In Denmark, women were 20% less likely to use ChatGPT than men in the same occupation, and 16% less likely within the same firm and while performing the same tasks. Women non-users more often cited a need for training (Humlum and Vestergaard, 2025). Since AI tools can raise productivity, this gap risks feeding future gaps in pay and progression, especially as unpaid care work limits women's time for reskilling (Fernandez et al., 2026).

### 3.1.2. Gender bias from recruitment to workplace

Although AI is often promoted as a way to make recruitment fairer, it can reproduce, and even amplify, the biases it is meant to remove, including gender bias (Rigotti and Fosch-Villaronga, 2024). Systems trained on past hiring decisions can perpetuate historical discrimination, while assumptions about gender, race, or disability embedded in their design can distort how candidates are assessed (EIGE, 2021). Those using these technologies, however, often underestimate this risk. A study for Belgium's Institute for the Equality of Women and Men, surveying more than 400 recruitment professionals, reported that about three-quarters used AI at some stage of recruitment, yet gender bias was seldom mentioned spontaneously as a risk (Wuidar et al., 2026).

This bias can enter at every stage of recruitment. At the advertising stage, science, technology, engineering, and mathematics (STEM) job ads have been delivered disproportionately to men even when their content was gender-neutral (Lambrecht and Tucker, 2019). At the screening stage, ChatGPT-based CV screening has been found to favour male-coded attributes and to be influenced by gendered names (Sivakaminathan and Musi, 2026). By contrast, Veldanda et al. (2023), replicating a classic correspondence experiment on names (Bertrand and Mullainathan, 2004) with GPT-3.5, Bard, Claude, and Llama, found the models robust to gender and race, but inconsistent on pregnancy status, a form of bias that affects women specifically. Assessment tools are also affected: in Germany, a Bayerischer Rundfunk investigation and Report München (2021) found that that the personality scores produced by a video-interview AI changed in response to features unrelated to a candidate's personality, such as wearing glasses or a headscarf, or the lighting of the video, including when a Black woman's video was darkened. Although the investigation did not show that these changes reduced candidates' chances of employment, it showed that these technologies react to appearance and recording conditions in ways that could disadvantage some groups, including women who wear religious attire.

Once in work, women can be managed by AI systems, although men are somewhat more likely to be subject to algorithmic management. In a worker survey conducted for the EC, 55.5% of male workers reported that algorithmic management tools are used in their workplace very rarely, compared with 49.3% of female workers, a gap that holds across all managerial functions (Paliokaitė et al., 2025). Nonetheless, lower average exposure does not mean lower risk, as algorithmic management is likely to have gender-specific impacts given existing inequalities in paid and unpaid work (EIGE, 2021). Algorithmic scheduling, which is common in female-dominated sectors such as retail and hospitality, allocates shifts at short notice, increasing income insecurity and stress, and potentially making paid work harder to combine with care. Real-time monitoring can penalise teleworkers interrupted by children or family members, who are disproportionately women. Women are also a growing share of platform workers, and slightly more likely than men to report unfair treatment based on their gender (EIGE, 2021). In 2020, the Court of Bologna ruled that Deliveroo's rider-ranking algorithm was discriminatory because it penalised absences without considering their reasons, such as a child's illness (Aloisi and De Stefano, 2021; Tribunale di Bologna, 2020). Although the case did not turn on gender, it shows how seemingly neutral metrics disadvantage those with caring responsibilities.

### 3.1.3. Intersecting inequalities in the automated workplace

As the previous sections show, AI systems can cause specific harms where gender intersects with other personal characteristics and social inequalities. The video-interview case investigated by Bayerischer Rundfunk and Report München (2021) exemplifies the potential for these technologies to respond to personal characteristics at the intersection of gender, race, and religion. The intersection of gender with care responsibilities is also widely attested (Aloisi and De Stefano, 2021; EIGE, 2021). The above-mentioned algorithm of the Austrian Public Employment Service equally made this intersection explicit, by deducting points for care obligations from mothers but not fathers (Knecht and Pühringer, 2018). A similar pattern emerged in Poland, where public employment offices used an algorithmic tool between 2014 and 2019 to sort unemployed people into three profiles, which determined their obligations and the support they could receive. A qualitative study of two offices found that staff routinely placed 'mothers of young children' in the profile with the fewest obligations but also the least support, thereby excluding them from activation measures, while fathers were not regarded as having childcare responsibilities (Palęcka and Sztandar-Sztanderska, 2024).

AI also holds potential for women with disabilities, for instance as assistive technology that opens access to employment. However, access to such technology remains deeply unequal (Hoekstra, 2026; Lulli et al., 2024). Finally, most labour market studies disaggregate data by gender alone, which results in the most affected women also being the least visible in policy design.

## 3.2. Public and private services

As public authorities and private companies increasingly rely on AI to make their services more efficient, AI systems can shape access to essential services such as welfare benefits, credit, and insurance. In the public sector, the 'digital welfare state' uses AI systems for tasks ranging from eligibility assessment and benefit calculation to fraud detection and risk scoring (Kaun and Liminga, 2025). In the private sector, banks and insurers use them to assess creditworthiness and price insurance (European Banking Authority, 2025; European Insurance and Occupational Pension Authority, 2024).

In both settings, these AI systems sort people into categories that carry moral and economic weight, such as 'deserving' and 'undeserving' welfare recipients (Huyskes, 2026), and can reproduce the gendered assumptions behind them. The distinction between public and private use matters, as public authorities exercise coercive power, although the line is blurring as public bodies sometimes partially outsource technological development to private companies (Geiger et al., 2023).

### 3.2.1. Welfare and social security

Risk-scoring systems designed to detect fraud and errors are now used in welfare systems across Europe (Geiger et al., 2023), including in schemes that mainly affect women, such as childcare and parental benefits. Most of these technologies work in a similar way. They combine personal data held by public authorities, such as income, employment, address, or family situation, to calculate a 'risk score' for each person receiving benefits. Those with the highest scores are then selected for investigation. Some systems follow rules set in advance by officials; others use ML, meaning that they learn which characteristics to treat as suspicious from past cases. Either way, people are primarily flagged because they resemble someone who was investigated before.

The Netherlands offers two well-documented examples. In the so-called 'childcare benefits scandal', the Dutch Tax Administration used a risk-classification model from 2013 onwards to select childcare benefit applications for checks. The model learned from earlier cases labelled as fraud, and even minor mistakes in applications were treated as grounds to demand that families repay all the benefits they had received, forcing thousands of families into debt. Subsequent investigations revealed that the

model assigned higher risk scores to people with dual nationality, those of Caribbean origin, and single mothers. These criteria were not chosen at random: they reflect a longer history in which ideas of who deserves welfare in the Netherlands have been shaped by race and gender (Huyskes, 2026). The gender dimension is even clearer in Rotterdam's welfare risk-scoring system. Between 2017 and 2021, the municipality used a ML model to rank every welfare recipient in the city by their estimated risk of fraud, and referred the top 10% for investigation. Journalists who obtained and rebuilt the model found that being a woman, being younger than 30, having children and speaking little Dutch were among the characteristics that most increased a person's risk score. These characteristics also added up: a woman supporting her children scored higher than a man in the same situation, and higher still if she had a migrant background (Van Schie et al., 2025).

Similar patterns have emerged elsewhere. As mentioned in Chapter 2, in Sweden, an analysis of over 6,000 people flagged by the Social Insurance Agency's algorithm for the temporary parental benefit for sick children found that it disproportionately flagged women, migrants, low-income earners, and people without a university degree (Granberg et al., 2024). In France, the family allowance fund (CNAF) has used a risk-scoring algorithm since 2010 whose 41 criteria, including low income, unemployment, and receipt of disability benefits, lead controls to target lone-parent families, most of them headed by women. A coalition of 15 organisations challenged this before the Council of State in 2024 (Amnesty International, 2024b). In Denmark, Amnesty International (Amnesty International, 2024a) found that the automation of the benefits system enables mass surveillance and creates barriers to access for marginalised groups, including women in crisis centres.

### 3.2.2. Credit scoring, lending, and insurance

In the private sector, AI increasingly shapes who can borrow, on what terms and at what price people can insure themselves. These systems are deployed in an unequal financial landscape, since women's working lives, and therefore their financial histories, tend to differ from men's. In 2024, women in the EU were far more likely than men to work part-time (27.8% compared with 7.7%), and also more likely to hold temporary contracts (11.3% compared with 8.9%) or to be underemployed (3.6% compared with 1.6%) (Eurostat, 2026). As a result, AI systems that reward high income, continuous employment, and long financial histories can translate these inequalities into lower scores or higher prices for women.

Removing gender from credit-scoring models does not, however, make them 'neutral'. Andreeva and Matuszyk (2019) analysed car loans granted to 79,386 customers of a major European bank between 2003 and 2009. Women had a lower probability of default than men. However, when gender was removed from the scoring model, women's chances of being accepted for credit fell while men's rose, without any gain in the model's predictive accuracy. Other variables in the model acted as proxies for gender, and women made up only a quarter of the sample, which may have amplified the bias. Kelley et al. (2022) reached a similar conclusion using data from a fintech lender: prohibiting the use of gender data substantially increased discrimination against women.

Evidence on insurance is more limited. Since the Court of Justice's *Test-Achats* ruling in 2011, insurers in the EU may no longer charge women and men different premiums on the basis of their sex (EU CoJ C-236/09, 2011; Hacker, 2018). The main risk is that AI reintroduces gender indirectly, through personal characteristics closely tied to it. The EC's guidelines (EC, 2012) on the ruling illustrate the distinction: in motor insurance, a car's engine power may be used because it affects accident risk, whereas a driver's height and weight may not, since they matter only as stand-ins for gender. The ML methods widely used in insurance pricing offer no guarantee against such indirect discrimination (Simon et al., 2026). Nonetheless, unlike in credit, there is so far little published evidence of how this affects women in

practice. Insurers' pricing models are proprietary, and detecting proxy discrimination generally requires data on policyholders' gender, which insurers are not allowed to use for pricing (Lindholm et al., 2022).

## 3.3. Healthcare

Healthcare is already shaped by sex and gender. Both biological and socio-cultural characteristics affect risk factors, prevalence, symptoms, prognosis, and treatment effectiveness across many conditions (Cirillo et al., 2020). Nonetheless, medicine has long treated the male body as the norm, and women remain under-represented in medical research (Merone et al., 2022). This has serious consequences for women's health, as two examples illustrate. First, because drug trials have largely neglected sex differences in how the body processes medication, women risk being overmedicated and experience adverse drug reactions nearly twice as often as men (Zucker and Prendergast, 2020). Second, women tend to be diagnosed later than men. When analysing hospital data for 6.9 million people in Denmark between 1994 and 2015, Westergaard et al. (2019) found that women were diagnosed later than men for 770 diseases, on average 2.5 years later for cancer and around 4.5 years later for metabolic diseases like diabetes.

AI is now being introduced into healthcare for different purposes, including diagnosis and treatment recommendations, patient monitoring and engagement, and administrative tasks (Lau, 2024). The following subsections examine how gender bias in AI affects general and reproductive healthcare, before turning to the specific case of femtech.

### 3.3.1. Gender bias in clinical and administrative uses of AI

The gender gap in medicine is reproduced in AI. Health data is often unstructured, inaccurate and unrepresentative (Lau, 2024), and developers seldom disclose the demographic composition of the data on which their models are trained.. A systematic review found that a significant share of published clinical AI models do not report the gender or racial distribution of their training data, raising doubts about whether they work equally well for everyone (Otokiti et al., 2025). More generally, research has begun to document gender bias across most uses of AI in healthcare.

In diagnosis, imbalanced data lowers performance for those who are under-represented. Larrazabal et al. (2020) demonstrated this with chest X-rays: across three deep learning models and two public datasets, diagnostic accuracy consistently dropped for the under-represented gender whenever the training data fell below a minimum balance. Bias can also stem from data that reflects unequal care. Four published models that predict liver disease from blood tests missed 44% of cases among women but only 23% among men, according to a study that recreated them (Straw and Wu, 2022). None of the 30 studies they reviewed had examined sex differences.

When it comes to treatment recommendations and patient engagement, LLMs raise new concerns. When asked whether patients should manage their condition at home, come in for a visit, or receive further tests, four LLMs, including GPT-4, made about 7% more errors for female patients and were more likely to advise women to self-manage at home, even when all gender cues had been removed from the clinical notes (Gourabathina et al., 2025).

AI and other prediction tools are also used in care specific to women's bodies, such as obstetrics, fertility treatment, and breast cancer screening. In these contexts, the risk is that AI systems are built around a narrow image of women, leaving out those who do not fit it. In other words, because only women are concerned, bias here takes the form of disparities between women rather than between women and men, which the intersectional definition adopted in this study captures as gender bias. In obstetrics, for example, a widely used calculator predicting the chance of a successful vaginal birth after a previous caesarean gave Black and Hispanic patients lower scores than white patients with

similar clinical profiles, potentially steering them towards repeat caesareans (Grobman et al., 2021). Although the calculator is a statistical model rather than AI, it shows how prediction tools in sex-specific care can encode intersecting biases. In fertility treatment, a randomised trial across 14 clinics in Australia and Europe could not show that a deep learning tool selected embryos for IVF as well as trained embryologists, although it was ten times faster (Illingworth et al., 2024). Breast cancer screening offers a more positive picture. Within Sweden's national programme, the Mammography Screening with Artificial Intelligence (MASAI) trial found that AI-supported mammography detected 20% more cancers without increasing false positives, while reducing radiologists' workload by 44% (Lång et al., 2023). Evidence from Hungary points in the same direction. In a prospective study at four screening sites, using an AI tool as an additional reader of the mammograms of 25,065 women raised the cancer detection rate by between 5% and 13%, depending on the phase, with only a marginal increase in recalls (Ng et al., 2023).

Finally, administrative tasks are not neutral either. In English adult social care, where over half of local authorities use LLMs to support social workers, a study of gender-swapped case notes for 617 older people found that Google's Gemma described men's physical and mental health needs more directly and downplayed women's more often, frequently describing women as managing well 'despite' their impairments. Meta's Llama 3 showed no such difference (Rickman, 2025). Because access to care depends on need assessments, these summaries could leave women with less support.

### 3.3.2. The case of femtech

The term 'femtech' refers to software and devices designed to support women's health, from period-tracking and digital contraception apps to smart breast pumps and menopause wearables. These products promise women greater knowledge of, and control over, their bodies (McMillan, 2024). However, they raise several concerns. Because they process highly sensitive data, they pose significant risks to privacy and data protection (Siapka and Biasin, 2021). They also give rise to a specific form of gender bias: the homogenisation of womanhood, which reduces the diversity of women's bodies and lived experiences to a single norm.

For example, many fertility apps rely on a small set of factors derived from averages or a preconceived norm, such as the 28-day menstrual cycle (Wachter-Boettcher, 2018). However, this 'textbook' cycle is the exception rather than the rule. In an analysis of more than 600,000 cycles from around 125,000 users in Sweden, the United Kingdom and the United States, only 13% lasted 28 days, and the average was 29.3 days (Bull et al., 2019). Apps nonetheless tend to assume that users not only have a regular cycle, but also track it for fertility purposes, identify as female and have a single male partner (Fox and Epstein, 2020).

This norm weighs most heavily on those whose gender intersects with other personal characteristics. For example, apps that require users to identify as female and adopt stereotypically feminine designs exclude trans men and non-binary people who menstruate, while the assumption of a male partner excludes women in same-sex relationships. In a study of around 2,000 app reviews and a survey of 687 users, pink, flowery designs, and the assumption of a female user with a male partner were among the main complaints (Epstein et al., 2017). These complaints may seem to concern user experience rather than gender equality. However, when measured against the standard of substantive equality adopted in Section 2.3, the design choices behind them are problematic in two ways. First, they reproduce stereotypes about who menstruates and in what kind of relationship, which substantive equality seeks to address. Second, where these assumptions are also built into an app's functionality, for instance through gender options that misgender users or predictions structured around a heterosexual fertility goal, they can make the app less useful to those who do not fit the norm, or deter them from using it

altogether. This disadvantages them in access to health-related information on the basis of their gender identity or sexual orientation. Evidence on whether femtech performs differently by race or ethnicity, by contrast, still appears to be lacking.

Homogenisation also undermines accuracy, since predictions are only as reliable as the assumptions behind them. They are least reliable for people with irregular cycles, for instance due to polycystic ovary syndrome (Corbin, 2020; Hendl and Jansky, 2022). Furthermore, few apps seem to be grounded in medical evidence: just 5% of contraceptive apps draw on medical literature (Earle et al., 2021), and those based on the calendar method have typical-use pregnancy rates of around 6% to 9% (Duane et al., 2016). Even Natural Cycles, the first app certified as a contraceptive in the EU, used to have a stated typical-use effectiveness of 93% in 2018 (Swedish Medical Products Agency, 2018), showing a gap between a technological promise and its real-world reliability.

## 3.4. Social media

This section examines social media as spaces of everyday social life, where people connect, present themselves, seek information and entertainment, and increasingly earn a living. The AI systems behind these platforms determine how content is organised, made visible, and personalised. They operate in a social context that is already deeply gendered: social media reflect and shape gender stereotypes, normative beauty standards, and expectations around women's sexuality. Online platforms usually require users to state their gender at sign-up, especially to enable targeted advertising, and can also infer it from behaviour (Fosch-Villaronga et al., 2021). The following subsections examine three AI systems through which gender bias operates on social media, namely recommender systems, content moderation, and advertising systems.

### 3.4.1. Recommender systems

Recommender systems determine what users see, watch, and listen to every day. In doing so, they have taken over a role long played by magazines and television in shaping gender norms, such as beauty ideals (Wohlfarth et al., 2026). The femininity that circulates on social media appears to be youthful, slim, white, non-disabled, and cis-heteronormative, thereby exclusionary (Gill, 2021). Recommender systems amplify these images. Instagram's Explore feed strongly reinforces beauty ideals once users routinely like or save such content, but adapts far less to content that does not reproduce them (Wohlfarth et al., 2026). In Europe, an investigation by AlgorithmWatch and the European Data Journalism Network tracked the feeds of 26 volunteers following 37 professional creators: pictures of women in bikinis or underwear were 54% more likely to appear, and pictures of bare-chested men 28% more likely, pushing creators, particularly women, to show more skin to reach their audiences (Kayser-Bril et al., 2020). Beauty filters compound the effect by reshaping faces to conform with Eurocentric beauty standards (Riccio et al., 2024). Therefore, this gender bias has an intersectional nature: by treating whiteness as the norm of beauty, filters implicitly mark the features of non-white women as in need of correction.

The effects of recommender systems on well-being are difficult to measure. Harms may be psychological, emotional, or economic. They often go unnoticed by users and interact with offline factors (Park et al., 2023). Nonetheless, some negative consequences for women are documented. A systematic review of experimental studies finds that the consumption of idealised images on social media increases body dissatisfaction among young women, largely through appearance comparison (Fioravanti et al., 2022). Visual platforms such as Instagram are more harmful to body image than text-based ones, and editing selfies worsens the effect (Vandenbosch et al., 2022).

Beyond body image, the gender bias of recommender systems is structurally tied to gender roles and existing inequalities in visibility and recognition. Music streaming provides an example. Women artists already face discrimination in the music industry: they are under-represented in charts and award nominations and receive less radio airtime. On streaming platforms, they cluster at the lower levels of popularity, and common recommendation methods propagate the gender bias present in their training data (Ferraro et al., 2024). Lower algorithmic visibility can mean smaller audiences and lower earnings, reproducing in the digital economy the barriers women creators already face offline.

### 3.4.2. Content moderation

Content moderation is the practice through which social media platforms screen, rank, restrict, or remove user content, combining community guidelines, automated classifiers, and human reviewers (Gillespie, 2018). Gerrard and Thornham (2020) describe these elements as 'sexist assemblages' that reproduce normative femininities and police content related to women's bodies, in part because the AI systems involved reduce gender identities to over-simplified categories.

This gender bias is visible in how online platforms moderate women's bodies. Meta's adult nudity policy prohibits images of nudity, with limited exceptions such as breastfeeding, while male nipples are permitted (Bouchaud et al., 2024). Beyond outright removal, Instagram also limits the visibility of what it considers 'borderline' content through what users call the 'shadowban': posts are hidden from the 'explore page' without notification, reducing their reach. Drawing on her experience as a pole dancer and instructor, Are (2022) shows that this indirect and hidden form of censorship particularly affects posts depicting women's bodies, nudity, and sexuality.

The same double standard extends to sexual and reproductive health. A report by the Center for Intimacy Justice (2026) finds that 84% of women's health businesses and 76% of nonprofits surveyed had ads rejected by Meta, and 63% had organic content removed, while overtly sexualised ads for men's sexual health products were approved. Similarly, a white paper by CensHERship and The Case For Her (2025) documents the systematic removal, restriction, or down-ranking of medically accurate, non-sexual content on menstruation, menopause, fertility, postpartum care and sexual well-being, which is often labelled as 'adult content' or 'inappropriate nudity'. Of the women's health creators surveyed, 95% had experienced censorship and more than half self-censored to avoid takedowns. The consequences are concrete: women lose access to health information, and creators lose visibility, audiences, and income.

### 3.4.3. Advertising systems

Social media platforms act as intermediaries for advertisers, who can target specific demographics. Gender bias appears in who sees them, as shown for job advertisements in Section 3.1.2 in what ads show, and in what ads are allowed to show. On the content side, stereotyping in YouTube advertising draws mainly on physical characteristics such as body shape and height, while stereotyping based on gender roles is similar to that in traditional advertising (Roth-Cohen et al., 2023). What ads are allowed show depends on how online platforms review them, and here the double standard seen in content moderation reappears. Using the ad repository mandated by the Digital Services Act (DSA), AI Forensics (2024) identified over 3,000 ads run in France, Germany, Spain, Italy, and Poland that contained fully explicit nudity and sexual activity, mostly promoting sexual enhancement products for men. Meta had reviewed and approved them, and they reached over 8 million users in the EU, overwhelmingly men. Ads concerning women's health, by contrast, are frequently rejected: 84% of women's health businesses surveyed by the Center for Intimacy Justice (2026) reported that Meta had rejected their ads, while overtly sexualised ads for men's sexual health products were approved (see Section 3.4.2). Taken together, these findings suggest that online platforms apply different standards

to advertising depending on whether it concerns men's or women's bodies and health. AI Forensics (2024) also found that Instagram and Facebook removed the same explicit images within minutes when they were uploaded as ordinary posts, suggesting that paid content is subject to weaker scrutiny.

## 3.5. Other domains

The areas analysed above are not exhaustive. AI systems with gendered effects are used in many other domains, and two examples illustrate how far the problem extends.

The first concerns women as victims. Since 2007, the Spanish security forces have used VioGén, an algorithm that assesses the risk faced by women who report gender-based violence and determines the protection they receive. An independent audit found that the system gives too little weight to psychological and other non-physical forms of violence, so that only 3% of victims receive a risk score high enough to trigger effective police protection. Women without children were also systematically scored as lower risk than mothers (Eticas Foundation, 2022). A *New York Times* investigation found that at least 55 of the 247 women killed after being assessed by VioGén since 2007 had been rated as facing negligible or low risk (Satariano and Toll Pifarré, 2024).

The second example concerns women as travellers and migrants. Between 2018 and 2019, the EU-funded iBorderCtrl project piloted an AI 'lie detector' for travellers entering the EU through Hungary, Latvia, and Greece. Its deception detection had been tested on 32 people, with fewer women than men (Gallagher and Jona, 2019), and critics have argued that such systems reinforce gendered, sexualised, and racialised forms of border control (Hall and Clapton, 2021). More broadly, facial recognition systems tend to perform worse for women: false matches are consistently higher for women, and when lower-quality border photos are used, systems fail more often to match women to their own images (Albiero et al., 2022). Accordingly, women may be more likely to be wrongly stopped or subjected to additional checks.

In conclusion, across its different manifestations, gender bias in AI disadvantages women, and some groups of women in particular. This has negative consequences that range from economic and health-related to psychological and representational. Measured against the standard of substantive equality adopted in this study and defined in Section 2.3, this disadvantage amounts to gender-based discrimination. Whether it also constitutes gender-based discrimination in the legal sense, however, depends on the scope and categories of non-discrimination law across the EU and on how far these can capture algorithmic forms of discrimination, as examined in Chapter 5.

Chapter 4 focuses on TFGBV, which disproportionately targets or affects women because of their gender, thus constituting discrimination against women. As set out in Section 2.3, gender bias and TFGBV are mutually reinforcing obstacles to substantive equality.

# 4. TECHNOLOGY-FACILITATED GENDER-BASED VIOLENCE: A PERVASIVE AND EVOLVING PHENOMENON

KEY FINDINGS

- There is no agreed term or definition for technology-facilitated gender-based violence (TFGBV), which makes it harder for victims to name their experience, for researchers and services to compare evidence, and for states to design a coherent response, although overly rigid definitions would quickly be outpaced by new technologies.
- TFGBV takes many forms, and new ones keep emerging. They range from image-based sexual abuse, sextortion, harassment and stalking to abuse in the metaverse and simulated abuse with AI chatbots and sex robots. These frequently overlap.
- AI has made sexual digital forgeries easier to produce and has made them a business. The practice pre-dates AI, but a growing chain of tools and platforms, some available through mainstream app stores, now lets anyone create and distribute such material, and some providers profit from it.
- TFGBV is fundamentally gendered. Misogyny is the most prevalent form of online hate in the EU countries studied, and women are the overwhelming majority of those targeted by sexual deepfakes.
- Gender intersects with other characteristics to shape who is targeted and how. Women in public roles, Black women and women with disabilities face particular risks. In some cases, assistive technologies meant to support women's independence have been turned into tools of surveillance and control.
- Online and offline abuse form a single continuum. TFGBV often extends into victims' physical lives and is frequently one tactic within a wider pattern of intimate partner violence and coercive control.
- The harms extend from individuals to society. Victims report anxiety, depression and self-harm, financial costs and withdrawal from work and public life. Because abusive content can resurface indefinitely, victimisation may have no clear end, and the silencing of women erodes public debate.
- For the first time, EU-wide data exist on the technology-facilitated dimension of violence against women, but they cannot yet show trends. The 2026 EU survey found that 7% of women who have ever worked had experienced cyberharassment at work, and that 10% of women who have ever had a partner had been monitored by that partner online. Differences in method mean these figures cannot be compared with the 2014 survey.
- TFGBV is widely underreported. Victims often do not recognise the abuse or fear being blamed, and police responses are frequently inadequate.

While TFGBV is increasingly recognised as a pervasive and pressing dimension of gender-based violence, there is still disagreement over what to call it, what it encompasses, how it is experienced, and how extensively it has been measured. As set out in Section 2.3, TFGBV, like gender bias, undermines substantive equality, and the two reinforce each other. This chapter addresses these questions in turn: terminology (Section 4.1), manifestations (Section 4.2), common characteristics and impact (Section 4.3), and scale and data collection (Section 4.4).

## 4.1. Terminology

The earliest reference to TFGBV can be traced back to the United Nations (UN) Special Rapporteur on violence against women, its causes and consequences, Dubravka Šimonović, whose report to the

Human Rights Council in 2018 defined it as 'any act of gender-based violence against women that is committed, assisted or aggravated in part or fully by the use of ICT, such as mobile phones and smartphones, the Internet, social media platforms or email, against a woman because she is a woman, or affects women disproportionately' (OHCHR, 2018, para. 23).

At the time of writing, however, two terms seem dominant at the UN level. On the one hand, the United Nations Population Fund seems to sponsor the adoption of TFGBV, understood as 'an act of violence perpetrated by one or more individuals that is committed, assisted, aggravated and amplified in part or fully by the use of ICT or digital media, against a person on the basis of their gender' (UNFPA, 2021, p. 10). On the other hand, UN Women, with the support of the World Health Organization, has converged on the formulation of technology-facilitated violence against women as 'any act that is committed, assisted, aggravated, or amplified by the use of information communication technologies or other digital tools, that results in or is likely to result in physical, sexual, psychological, social, political, or economic harm, or other infringements of rights and freedoms' (UN Women and WHO, 2023, p. 4). Read together, the two definitions share a common structural logic: an act or omission that is mediated by technology, targeting a person on the basis of gender. Where they diverge is in scope and precision. The definition of UN Women is explicitly framed around women as the object of protection and is more detailed in describing the harm, while the definition of UNFPA is gender-neutral in its wording.

The influence of international organisations on the harmonisation of legal and policy definitions of TFGBV is increasing. Within Europe, and within its mandate to offer guidance on the interpretation of the Council of Europe (COE) Convention on Preventing and Combating Violence against Women and Domestic Violence (the Istanbul Convention), the Group of Experts on Action against Violence against Women and Domestic Violence (GREVIO) has adopted the term 'the digital dimension of violence against women', defining it as covering 'both online aspects (activities performed and data available on the internet, including internet intermediaries on the surface web as well as the dark web) and technology-facilitated (activities carried out with the use of technology and communication equipment, including hardware and software) harmful behaviour perpetrated against women and girls' (GREVIO, 2021, p. 23).

The EU, for its part, has settled on 'cyberviolence', likely to preserve consistency with the EU's existing legal vocabulary on cybercrime. EIGE defines it as covering 'a range of different forms of violence perpetrated by ICT means on the grounds of gender or a combination of gender and other factors (e.g. race, age, disability, sexuality, profession or personal beliefs)' (EIGE, 2022, p. 4). Directive (EU) 2024/1385 on combatting violence against women and domestic violence follows the same terminological preference but fails to provide a standalone definition, referring instead, in general terms, to the capacity of ICT tools to amplify violence in ways that cause profound and long-lasting harm in its Recital 17.

Within academic literature, terminological variety is even broader, and the choice of term may signal the author's disciplinary background or jurisdiction as much as it describes the abuse itself. Amongst the labels used are 'technology-facilitated abuse' (Koukopoulos et al., 2026), 'technology-facilitated sexual violence' (Powell, 2022), 'online violence against women' (Barker and Jurasz, 2020), 'cyberviolence' (Backe et al., 2018), 'cyber gender abuse' (Citron, 2023), 'online victimisation' (Gámez-Guadix et al., 2015), and 'technology-facilitated coercive control' (Dragiewicz et al., 2018). It has also been suggested that an umbrella definition be replaced with a four-fold classification based on technology's role in channelling the gendered harm: technologies that facilitate the commission of gender-based violence; ICTs that aggravate or amplify the harm victims experience; technologies that give rise to entirely new forms of abuse; and cyberspace itself, used as a vector to enable gender-based violence (Rigotti and McGlynn, 2025).

This definitional fragmentation has practical consequences. First, it makes it difficult for victims to name and articulate their own experience, particularly where the abuse falls between categories or where no single term captures its full character. Second, it complicates the work of stakeholders, such as researchers, service providers, and Civil Society Organisations (CSOs), who need to understand, compare, and evaluate the abuse across contexts. Third, it constrains the ability of states to design an adequate legal and policy response, especially given the inherently borderless nature of TFGBV (Rigotti and McGlynn, 2025). Nonetheless, rigid definitions risk freezing an understanding of the phenomenon that will inevitably be overtaken by new technological forms of abuse. Therefore, there is tension between the need for flexibility that keeps pace with victims/survivors' lived experience and the need for enough definitional stability to support consistent legal and policy action. Any terminology adopted must also guard against framings that risk victim-blaming or otherwise misleading characterisations of the abuse.

Against this background, this study adopts the term 'technology-facilitated gender-based violence'. It is considered the most comprehensive and inclusive of the available options, both because it is technology-neutral and future-proof, and because it extends to gender in all its identities and expressions.

## 4.2. Manifestations

The overview that follows is not exhaustive. Manifestations of TFGBV evolve as new technologies are developed and used, and they frequently overlap. Terminology, too, remains unsettled: terms are often revised or displaced, not least by the jargon of online communities themselves. Image-based sexual abuse.

### 4.2.1. Image-based sexual abuse

The term 'image-based sexual abuse' refers to 'all forms of the non-consensual creating, taking or sharing of intimate images or videos, including altered or manipulated media, and threats to distribute such material' (Rigotti and McGlynn, 2022, p. 454). It was coined to challenge 'revenge pornography', a label that implies victims were responsible for the abuse they suffered and overlooks the power dynamics often at play (McGlynn and Rackley, 2017). The term is also misleading since these acts are frequently produced for purposes other than sexual arousal, and go beyond a straightforwardly 'pornographic' narrative (Semenzin and Bainotti, 2020). Finally, 'revenge pornography' fails to capture the full spectrum of the issue, only focusing on the malicious distribution, by former partners, of sexual and intimate images obtained with the victim's original consent (McGlynn and Rackley, 2017). At the EU level, the term "non-consensual intimate imagery" (NCII) is also increasingly used. It refers to a specific category of image-based sexual abuse, namely intimate content, whether authentic or synthetic, that is created, shared or disseminated without consent (Rigotti et al, 2026). However, the term 'image-based sexual abuse' remains the more comprehensive frame, encompassing manifestations such as:

- **Recordings in private and public settings**: typically via hidden cameras placed in spaces where a person has a reasonable expectation of privacy, such as changing rooms, bathrooms, or bedrooms (McGlynn and Rackley, 2017). More recently, this harm has increasingly been perpetrated through smart glasses and other wearable digital devices (Stilinovic et al., 2026).
- **Recording of sexual violence**: the filming or photographing of a sexual assault or rape as it occurs, often by the perpetrator, which can serve to intimidate or silence the victim and may later be distributed or used as a further tool of coercion (Dunn, 2020).

- **Upskirting**: the covert taking of a photograph or video beneath a person's clothing, without consent, to capture genitals, buttocks, or underwear that would not otherwise be visible, typically committed in public or crowded spaces (Lewis and Anitha, 2022).
- **Downblousing**: the analogous act of covertly photographing or filming down the front of a person's top or blouse, without consent, to capture their chest (Lewis and Anitha, 2022).
- **Screenshotting**: capturing, through screenshots or screen recordings, images or videos that were shared only temporarily or only with the recipient, for example through disappearing messages or during a live video call, in order to keep or distribute them without consent (Ringrose et al., 2022).
- **Cyberflashing**: the digital distribution of penis images or videos to another without the recipient's consent, including flaccid and erect penises, as well as altered and live-streaming images (McGlynn and Johnson, 2021, p. 3).
- **Sexual digital forgeries**: see definition in Section 4.2.2.
- **Sextortion**: see definition in Section 4.2.3.

Recent cases in the EU illustrate the range of these harmful practices. In August 2025, Meta removed "Mia moglie", an Italian Facebook group with more than 32,000 members in which men posted intimate images of their partners and other women without their consent (Reuters, 2025). Instant messaging services also play a key role. An ethnographic study of 50 Italian Telegram groups and channels found communities of tens of thousands of users exchanging non-consensual intimate images, hidden-camera footage and material taken from other social media platforms, where the lack of consent was treated as "one of the rules of the game" (Semenzin and Bainotti, 2020). In the Netherlands, an investigation by the public broadcaster NOS identified nine "exposure" groups on Telegram with more than 10,000 members each, the largest with around 85,000, generating more than 10,000 messages a day. About four in five of the messages analysed combined sexual or nude images with the victim's contact details. Women with a migration background were particularly targeted (Schellevis and van Hulzen, 2023).

One final consideration concerns the connotation of the term 'image-based sexual abuse'. Although named 'sexual', the material is understood more broadly, encompassing intimacy in an inclusive and culturally diverse way (Rigotti and McGlynn, 2022). This is evident in a report by the organisation Chayn, which shows how a photo without a hijab and a video of dancing at a wedding have been used to target women across the Pakistani diaspora, causing victims harm comparable to that arising from content of a sexual nature (Chayn, 2026).

### 4.2.2. Sexual digital forgeries

McGlynn and Toparlak (2025) define sexual digital forgeries as non-consensual sexual or intimate material generated or manipulated using AI. The underlying practice, however, pre-dates AI: manual editing tools such as Photoshop have served the same purpose for decades (Hawkins et al., 2025). What AI has changed is not the act of fabrication itself, but its scale and accessibility.

Popular usage continues to favour different labels, including 'deepfake pornography' and 'sexual deepfakes' (Dunn, 2024; Hawkins et al., 2025). Advances in AI have sharply lowered the technical skill required to produce convincing synthetic sexual imagery of identifiable people, and technology providers – along with, in some cases, the creators using their tools – have moved to monetise this newfound accessibility (Cui et al., 2026; Gómez Ortega et al., 2026; Qiwei et al., 2026). The scale of the

resulting harm was underscored by the Grok scandal of January 2026, in which xAI's chatbot was used to mass-produce non-consensual sexualised images of women and girls (Bouchaud, 2026).

Additionally, public debate has often centred on so-called 'nudification applications', namely tools that take a photograph of a clothed person and, typically without consent, generate a version of them nude or partially undressed (Williams, 2025). This framing, however, undersells the range of what the underlying technology allows: the same class of AI systems can alter body characteristics such as breast size or apparent age, apply stylistic filters, or insert the subject into explicit sexual scenes or videos, in some cases even simulating sexual violence (Gibson et al., 2024). At the same time, several distinct categories of tools contribute to the creation and dissemination of sexual digital forgeries, sitting at different points along a spectrum between raw technical capability and finished, circulated content. Resource-sharing platforms distribute the underlying components, such as open-source image-generation models. Ready-to-use web applications turn these into simple upload-and-generate services. Some are built specifically for nudification and have occasionally been available in mainstream app stores. Finished material then circulates through file-hosting services and private communities, for example on Telegram, while commissioning platforms let users pay creators for customised content (Cui et al., 2026; Daffalla et al., 2026; Williams, 2025).

Women in public life are frequent targets. In the Netherlands, the television presenter Welmoed Sijtsma discovered a sexual video in which her face had been superimposed onto the body of a porn performer (NL Times, 2023). A few months later, the Dutch daily *AD* found a large number of similar videos of Dutch public figures, including television presenters, Olympic athletes, current and former ministers, and mayors, on a platform with around 13 million monthly visitors (NOS, 2024b). In Germany, the actress and presenter Collien Fernandes alleged in 2026 that for years she had been impersonated on dating apps and pornography-sharing websites, through sexually explicit material sent to men in her name, an AI-generated version of her voice, and synthetic videos depicting her being raped (Scally, 2026).

### 4.2.3. Sextortion

Sextortion refers 'to making threats to share nude or sexual images to coerce the victim into complying with certain behavioral or financial demands' (Ray and Henry, 2025, p. 138). The term covers different dynamics, distinguished primarily by what the perpetrator is seeking to obtain, e.g. money, further content, sexual acts, or broader compliance within an ongoing pattern of control (Ray and Henry, 2025).

What sets sextortion apart from other forms of image-based sexual abuse is that offender and victim may never meet, or even interact, offline. Non-consensual content can be obtained through manipulation or coercion alone, with no prior physical relationship required. It is this subsequent use of the material as a threat that defines the harmful conduct, especially since what is central to sextortion is the power an offender holds over the victim through the mere possibility of harm (O'Malley and Holt, 2022). Cyberspace compounds this dynamic further, providing offenders with global reach to potential victims and relative anonymity (Henry and Umbach, 2024).

This is evident in a case from Greece. In 2024, a 61-year-old woman met a 51-year-old man online. Their communication took on a sexual character, and they exchanged photos and had a video call with sexual content. According to the woman, he then threatened her almost daily, through messages on social network or by phone, that he would send the images and video to her relatives and people close to her, and post them in large social media groups, unless she agreed to have sex with him (Aleiferis et al., 2024). The case also shows that TFGBV is not confined to young people, who tend to be the focus of public debate.

### 4.2.4. Cyberharassment

In the absence of consensus, cyberharassment is generally understood to cover a range of behaviours, conducted through ICTs and AI systems, that share the aim or effect of intimidating, distressing, or humiliating a victim (Blackwell et al., 2018; Citron, 2009; Schoenebeck et al., 2023). Manifestations include:

- **Threatening conduct**: repeated or continuous conduct directed at a person online that conveys a threat to commit criminal offences such as rape (Citron, 2009).
- **Coordinated 'pile-on' or 'networked' harassment**: threatening or insulting conduct jointly engaged in by multiple actors, whose cumulative effect produces harm that no single actor's contribution would independently cause (Citron, 2009; Dunn, 2020). This dynamic can be artificially amplified by networks of bot or fake accounts, which inflate the reach of the harassing content (Burgess, 2023), or generated directly by AI chatbots themselves (McGlynn et al., 2026).
- **Flaming**: the posting of hostile, insulting, or deliberately provocative messages in a public online space, typically as part of an escalating exchange between individuals (Schoenebeck et al., 2023).
- **Doxxing**: the non-consensual public disclosure of a person's personal data, intended to expose them to harm from others. Its consequences go beyond the immediate loss of privacy and anonymity to cascading secondary abuse, including death threats, extortion, and harm to the victim's family (Anderson, 2026).
- **Impersonation and identity theft**: the creation of fake profiles or placement of advertisements in a victim's name, often falsely suggesting sexual availability, or the unauthorised use of a victim's personal information to access and use their accounts or services (Dunn, 2020; Schoenebeck et al., 2023).

Recent cases show how coordinated harassment operates. The French "Ligue du LOL", revealed by Libération in February 2019, was a private Facebook group of around thirty journalists and online influencers who, over several years, coordinated attacks on other users of the social network, especially feminists, LGBTIQA* people and racialised people (Andraca, 2019). In Slovakia and Czechia, an analysis of more than 8,600 public posts on Facebook, Instagram, TikTok, YouTube, and Telegram between 2022 and 2025, which together reached over 16 million views, found that women politicians and journalists were targeted with sexualised insults, belittling language, claims that they had "slept their way to the top", and threats. Those targeted reported anxiety, fears for their safety and withdrawal from public life (Kužel et al., 2025).

### 4.2.5. Cyberstalking

Cyberstalking refers to the repeated or persistent use of ICTs and AI systems to monitor, track, or contact a person without their consent, in a manner that causes fear, distress, or a loss of sense of safety (Weekes et al., 2025). While the literature agrees that such behaviours must be unwanted by the victim, definitions differ in their requirements as to duration or frequency, perpetrator intent, and victim fear (Wilson et al., 2022).

Manifestations include GPS tracking of a victim's location; the covert installation of stalkerware or spyware on a victim's device; unauthorised access to a victim's accounts through password or identity theft; monitoring of a victim's social media activity and communications; persistent unwanted contact via messaging, email, social platforms, and dating apps, sometimes despite explicit requests to stop;

and the misuse of smart home or Internet of Things (IoT) devices to surveil a victim's movements within their own home (Ayerza et al., 2026; Wilson et al., 2022). Cyberstalking often functions as one tactic within a broader strategy of coercive control, rather than as an isolated act of surveillance (Dragiewicz et al., 2018). More recently, AI chatbots have been found to intensify this pattern, providing perpetrators with detailed and personalised guidance for tracking or contacting a victim (McGlynn et al., 2026).

An investigation by the German outlet netzpolitik.org and the public broadcaster SWR into a leak of 3.6 million customer-service messages from mSpy, a commercial stalkerware application, found that many users sought to secretly monitor their partners, with more than 42,000 messages in German alone (Köver and Schwarzbeck, 2025). One of the women the journalists spoke to had separated from a partner whom she had reported to the police for violence during their relationship. He kept appearing at her home, her workplace and along her routes, having secretly installed mSpy on her phone, which gave him access to her location, messages, passwords, photos, calendar and contacts. She developed post-traumatic stress disorder, depression, panic attacks and sleep problems, and withdrew from friends and family (Köver and Schwarzbeck, 2026).

### 4.2.6. Gender-based hate speech

Gender-based hate speech, sometimes termed 'sexist' or 'misogynistic', refers to the incitement to violence or hatred against a person or group of persons on the basis of gender, through the public dissemination of material via ICTs and AI systems (Bladini and Svensson, 2025). It reflects a well-documented and sharply rising phenomenon, reinforced online by the disinhibiting effect of presumed anonymity and a corresponding sense of impunity (Wagner and Carlson, 2025).

The identification of gender-based hate speech, however, raises interpretive difficulties, since much of it does not explicitly reference the target's gender, and biased motivation must instead be inferred from the content or context of dissemination. Sexualised threats and body-shaming comments are a classic example: gendered in substance even where not explicitly framed as such. This difficulty is compounded by the long-standing tension between hate speech and freedom of expression, which calls for a contextual resolution (Sjöholm, 2024). At the same time, it is sharpened by the social perception of gender-based hate speech as acceptable, even banal, since it is normalised within social institutions and everyday communication in ways other hate speech typically is not. Eurobarometer survey data demonstrate how far this normalisation extends: in Bulgaria (52%) and Lithuania (51%), a slim majority of respondents agree that women who share their opinions on social media should accept that they will receive sexist, demeaning or abusive replies (EC, 2024).

Monitoring data point to the same direction. Italy's Map of Intolerance, which tracks hateful posts on X, found misogyny to be the largest category of hostile content in 2025 (37%). According to its authors, the decline from 50% in 2024 signals normalisation rather than reduction: sexist insults have become generic, dehumanising terms in everyday language, making the underlying stereotype almost invisible and therefore more dangerous (Vox – Osservatorio Italiano sui Diritti, 2026). In Romania, researchers collected 2,022 comments on online news and Facebook posts about ten women in public life, mostly politicians, between November 2020 and January 2021, and classified 44% of them as sexist. The comments portrayed the women as sexually promiscuous or as having traded sex for their careers, mocked them for being childless, and dwelt on their bodies, clothes and hair colour (Moldovan et al., 2022).

### 4.2.7. Meta-rape

The metaverse has given rise to a further category of manifestations of TFGBV, named 'meta-rape' and defined as 'all forms of non-consensual intrusive conduct of a sexual nature in the metaverse, committed by or experienced through an avatar' (McGlynn and Rigotti, 2025, p. 562). In other words, the underlying premise is that forms of gender-based and sexual violence long documented offline can also occur within the metaverse: a computer-generated, three-dimensional space that users navigate via headsets and hand controls, interacting with one another and their environment through an avatar, or 'virtual self' (Ritterbusch and Teichmann, 2023). The literature has documented its occurrence in virtual worlds such as LambdaMoo, Second Life, and Horizon Worlds (Danaher, 2018; Patel, 2021; Strikwerda, 2015). In one widely reported case, Nina Patel described how a group of male avatars assaulted her avatar shortly after she entered Horizon Worlds, taking photos and commenting "don't pretend you didn't love it" (Patel, 2021).

### 4.2.8. Chatbot-simulated and sex robot-mediated abuse

Chatbot-simulated and sex robot-mediated abuse respectively refer to interactions in which a user engages an AI chatbot (e.g. Replika) or a sex robot (i.e. a human-like machine designed for emotional and sexual engagement, increasingly augmented by AI-powered capabilities), in the roleplay of sexual violence, with the technology itself standing in as the simulated victim (McGlynn et al., 2026; Rigotti, 2025). In most cases, the simulated victim is a fictional persona. However, chatbots and sex robots can also be designed or customised to reproduce the likeness of an identifiable person, such as a celebrity, an acquaintance or a former partner, whether through their name, appearance, voice or personality (Karaian, 2024; McGlynn et al., 2026). For example, in 2025, a Reuters investigation found that Meta's platforms hosted chatbots impersonating celebrities such as Taylor Swift, Scarlett Johansson, Anne Hathaway, and Selena Gomez without their consent (Horwitz, 2025).

Unlike the manifestations discussed above, this conduct involves neither a human being nor an avatar: the interaction takes place bilaterally between a single user and the AI system, whether embodied or not. Few safeguards currently exist to prevent this, and some systems are specifically designed to simulate a lack of consent (McGlynn et al., 2026; Rigotti, 2025). Despite the absence of a third-party human victim, this form of technology-facilitated abuse has been framed within the concept of cultural harm, which goes beyond traditional, individualised notions of harm to capture the broader socio-cultural effect of normalising and institutionalising the subordination and oppression of women (Vera-Gray, 2020). Where a system reproduces a real person's likeness, however, the conduct also affects an identifiable individual, bringing it closer to the forms of image-based abuse discussed above.

## 4.3. Nature and impact

Despite this diversity of manifestation, several characteristics of TFGBV recur across the literature. First, these harms are fundamentally gendered, disproportionately affecting women as a result of entrenched societal structures of inequality (Rigotti and McGlynn, 2025). Empirical research confirms this pattern. A content analysis of social media posts commissioned by the European Union Agency for Fundamental Rights (FRA) across Bulgaria, Germany, Italy, and Sweden found that misogyny was the most prevalent form of online hate: posts targeting women outnumbered those targeting people of African descent by nearly three to one, and most often took the form of abusive language and sexualised incitement to violence (FRA, 2023). Similarly, Home Security Heroes' 2023 *State of Deepfakes* report, based on an analysis of nearly 96,000 videos, found that content of a sexual nature accounts for 98% of all deepfake videos online, and that 99% of the individuals targeted within it are women (Home Security Heroes, 2023). In 2026, an investigation by AI Forensics into Telegram

communities in Italy and Spain reached a comparable conclusion, finding that women were not only the overwhelming majority of victims but were often named, tagged, and rendered locatable through shared profile links (Semenzin et al., 2026). This is not to discount the forms of gendered harm experienced by men, including sextortion, which likewise reflect rigid gender norms, while nonetheless remaining, in large part, perpetrated by men (Hayes et al., 2026; O'Malley et al., 2026).

Second, the risk of, and experience of, TFGBV depends on how a victim's gender intersects with other personal characteristics and structures of subordination. Public-facing women, such as journalists, activists, politicians, and athletes, are particularly targeted (Posetti et al., 2025). A World Athletics study tracking 161 athletes' social media accounts during the Tokyo 2020 Olympics found that 87% of all identified abusive posts were directed at women, with non-white female athletes receiving the greatest share of abuse (Wall Tweedie et al., 2026). This illustrates how gender and race compound within a single professional category, a pattern research shows extends more broadly (Amnesty International, 2018; Glitch UK, 2023), with more limited attention turned to the abusive experiences of the LGBTIQ+ community (Dunn, 2020). Attention has likewise turned to women with disabilities. Research into TFGBV of women with intellectual or cognitive disability showed perpetrators, sometimes the women's own carers, tampering with hearing aids, attaching GPS trackers to wheelchairs, and misusing smart home systems to control lighting and other devices (Woodlock and Harris, 2024). In other words, the assistive technologies were found to be repurposed as instruments of surveillance and coercive control.

Third, TFGBV occurs along a continuum of gendered and sexualised harms, which spans online and offline environments (Rigotti and McGlynn, 2025). Following Liz Kelly's theory of the continuum (1988), women's experiences of TFGBV resist a neat division into 'minor' and 'serious' incidents: everyday, normalised forms of online sexism and low-level harassment sit on the same continuum as more severe manifestations, such as image-based abuse (McGlynn et al., 2017). Furthermore, the continuum is spatial: technology-facilitated harm spills over into victims' 'physical' lives. This is evident in cyberstalking, where victims'/survivors' experience of domestic violence often involves surveillance moving between digital and physical spaces as part of a single pattern of abuse (Weekes et al., 2026). At the same time, TFGBV co-occurs with other forms of abuse. In 2026, for example, the Danish CSO Danner found that women who had experienced economic partner violence also experienced other, overlapping forms, including TFGBV (79%) (Edelenbos et al., 2026).

Lastly, a further characteristic common to TFGBV is its impact on victims and society at large. At the individual level, victims/survivors report psychological consequences including depression, anxiety, panic attacks, and self-harm (Bates, 2017; McGlynn et al., 2021; McGlynn and Johnson, 2021; O'Malley, 2023; Posetti et al., 2025; Schoenebeck et al., 2023). Survey evidence points to their extent. In a poll of 4,000 women in eight countries, including Denmark, Italy, Poland, Spain and Sweden, 23% had experienced online abuse or harassment. Of these, 55% reported stress, anxiety or panic attacks, 63% had trouble sleeping, 61% reported lower self-esteem or confidence, and 41% felt that their physical safety was threatened (Amnesty International, 2017). Among young adults in Italy who had experienced cyberstalking, most of them women, 58% reported mild to high levels of anxiety and 39% symptoms of depression (Begotti et al., 2022). In image-based sexual abuse, women often struggle with a perceived loss of control over their bodies and online appearance, and this is compounded by the persistence of abusive content in cyberspace: because the material can resurface, be rediscovered, or remain in circulation, victims/survivors face a form of re-victimisation with no clear endpoint (McGlynn et al., 2021). Interviews with 75 victim-survivors of image-based sexual abuse, who described it as "torture for the soul", capture this experience through five recurring harms: social rupture, constancy, existential threat, isolation and constrained liberty (McGlynn et al., 2021). Such harm does not depend on physical contact. Women interviewed about harassment in social virtual reality described feeling touch on their

physical bodies when their avatar was touched, freezing in fear, and experiencing shame and self-blame (Schulenberg et al., 2023).These harms are not restricted to mental health and well-being: victims sometimes miss work or school, bear significant costs to access legal or specialist support, and withdraw from social and digital participation, out of fear of further blame or repeated abuse (Paradiso et al., 2023).

At the societal level, online environments can become sites of oppression and marginalisation for women, manifested in their silencing, the erosion of public discourse, and the deepening of existing digital divides (Dunn, 2020; Posetti et al., 2025). Furthermore, the cumulative effects extend to measurable economic and public-health costs borne by society, a pattern already documented in relation to gender-based violence and estimated by EIGE to cost the EU €366 billion annually (EIGE, 2021). Estimates focused on the digital sphere confirm that these costs are considerable. Lomba et al. (2021) put the combined cost of cyberharassment and cyberstalking to individuals and society at between €49.0 and €89.3 billion. Most of this burden reflects harm to victims' well-being. Expressed in monetary terms, their reduced quality of life makes up roughly 60% of the cost of cyberharassment and 50% of that of cyberstalking. Consequences for victims' working lives form the second-largest share, at around 30% and 35% respectively, with the higher figure for cyberstalking attributed to reduced participation in the labour market. Healthcare and legal expenses account for the remainder.

## 4.4. Scale and data collection

In 2026, the FRA and EIGE released the *EU Gender-based Violence Survey – Evidence for Policy and Practice*, namely the first EU-wide survey to include dedicated questions on cyberharassment, alongside questions on intimate-partner psychological violence and stalking that also captured the technology-facilitated dimension. The results show that 7.0% of ever-working women in the EU have experienced cyberharassment at work. In the eight Member States where FRA and EIGE fielded an extended questionnaire, this rises to 11.7% for harassment outside work and 8.5% for cyberstalking, while 10.2% of ever-partnered women across the EU-wide report having been monitored by a partner via social media or location tracking as a form of controlling behaviour (FRA and EIGE, 2026).

A direct comparison with the previous EU-wide survey, however, is not straightforward, since it relied on different survey questions. In 2014, the survey measured cyberharassment using only two items, finding that 11% of EU women had been affected since the age of 15, and 5% in the preceding 12 months. Cyberstalking, measured separately, stood at 5% lifetime prevalence (FRA, 2014). Any claim of a trend between 2014 and 2026 would therefore overstate what the data can support. However, this comparability gap is what Directive (EU) 2024/1385 on combatting violence against women and domestic violence is designed, at least in part, to close. Article 44 obliges Member States to conduct regular population-based surveys, and EIGE has already developed a measurement framework for Member States to use to measure the statistical prevalence, incidence and reported cases of specific forms of TFGBV and to report the data collected in a comparable way (EIGE, 2025).

Other international studies, which nonetheless include EU respondents, are also of interest, not least because several disaggregate their data. Most recently, a UN Women-commissioned global survey conducted in partnership with UNESCO and the International Center for Journalists found that 'these forms of abuse, which were considered relatively rare just a few years ago, are now a significant portion of the broader ecosystem of online violence' (Posetti et al., 2025, p. 8). Disaggregated by professional role, 9% of women journalists and media workers participating in the survey reported non-consensual sharing of personal images. This rose to 15% among both human rights defenders/activists and writers and other public communicators. Earlier, in 2021, a study by the Economist Intelligence Unit found that

85% of women had personally experienced, knew someone who had experienced, or had witnessed TFGBV, a figure that fell to 74% among respondents in Europe (Economist Intelligence Unit, 2021).

At the national level of the Member States, there are also initiatives within civil society and academia to capture the scale of TFGBV. For example, in Germany, a study conducted by the Technical University of Munich in cooperation with the civil society organisation HateAid, surveying 1,114 politically engaged people across Germany, demonstrated that 58% had experienced digital violence, rising to 63% among women compared with 53% among men. Almost a quarter of affected women reported receiving threats of sexual violence, such as rape threats, nearly eight times more often than their male counterparts (3%). Nearly half of affected women said they had at least sometimes considered avoiding a political position especially exposed to digital violence, and 22% had considered withdrawing from political engagement altogether, compared with 10% of affected men (HateAid and Technische Universität München, 2025). In France, the Centre Hubertine Auclert focused on intimate partner cyberviolence, conducting a research-action study with 302 women receiving support for domestic violence, alongside the professionals accompanying them. Its central finding was that nearly nine in ten of these women reported having also experienced cyberviolence at the hands of their partner or former partner, most commonly cybercontrol (73%) and cyberharassment (63%), with three-quarters of affected women reporting two or more cumulative forms of cyberviolence (Centre Hubertine Auclert, 2018).

Overall, data play a key role in understanding TFGBV. Beyond establishing prevalence, as this section has sought to do, they reveal the forms abuse takes, who perpetrates it, how frequently it recurs, its consequences for victims/survivors, and whether its scale is changing over time. These are insights that have partially supported the analysis above and, together, can inform policy and service provision (Jansen, 2023). Nonetheless, one should not forget that evidence bases diverge considerably in their research objectives, methodologies, and methods, meaning that a comprehensive overview is nearly impossible to achieve (UNFPA and Wilson Center, 2022). This is also why the data reported in this study should be treated with caution, and their function understood as illustrative and indicative.

A further limitation, common to research on gender-based violence more broadly, and one the EU Gender-based Violence Survey (2026) documents, is underreporting. Victims may not recognise their own experience as a form of TFGBV in the first place. Many fear being blamed for what happened to them, and law enforcement agencies and legal professionals frequently respond with insufficient sensitivity, downplaying victims' experiences or offering little beyond generic advice, such as recommending stricter privacy settings (Allen and Hermolle, 2026; Dunn, 2021; Gkantona et al., 2026; Short et al., 2026). The fear of blame is well founded. In a 2024 Eurobarometer survey, 43% of EU respondents agreed that a woman who shares intimate pictures with someone is at least partly responsible if they are later shared online without her consent, with agreement highest in Latvia (74%) and Lithuania (70%) (EC, Directorate-General for Justice and Consumers, 2024).

# 5. LEGAL FRAMEWORK

KEY FINDINGS

- EU anti-discrimination law affords uneven protection across protected grounds, creating gaps in coverage and leaving (certain forms of) intersectional discrimination insufficiently protected. The continued deadlock on the Horizontal Equal Treatment Directive perpetuates structural asymmetries.
- The application of traditional discrimination concepts to algorithmic decision-making is challenging. Existing frameworks struggle to capture proxy discrimination, misclassification based on inferred characteristics, intersectional discrimination, and discrimination arising from data-driven grounds such as socio-economic status, health status or income, which largely fall outside current legislative categories.
- The AI Act introduces a risk-based governance framework incorporating non-discrimination safeguards. However, many AI systems fall outside its scope of application, as the Act's substantive obligations are primarily triggered by classification as a prohibited practice, a high-risk AI system, or a transparency-related use case. Compliance with the AI Act does not guarantee compliance with EU equality law, which applies independently of risk classification. Moreover, practices of regulatory 'de-risking' may remove harmful systems from the Act's scope.
- While gender constitutes a protected ground under equality law, it is not included among the special categories of personal data under Article 9(1) of the General Data Protection Regulation (GDPR).
- Although the DSA incorporates non-discrimination concerns through its reference to Article 21 of the Charter, its governance architecture remains centred on the mitigation of identified systemic risks and offers limited tools for addressing structural gender discrimination embedded in data, design and system architecture.
- Directive (EU) 2024/1385 establishes a common EU framework for addressing TFGBV and expressly acknowledges intersectional discrimination. However, the Directive does not define 'gender'.
- Recent legislative developments point towards a broader simplification agenda with potentially adverse consequences for accountability and redress. The withdrawal of the AI Liability Directive removes a dedicated avenue for remedies, the Digital Omnibus package postpones key AI Act obligations, and proposed amendments to the GDPR definition of personal data may reduce protection against discriminatory data practices.
- The effectiveness of existing safeguards remains constrained by structural implementation barriers, including persistent transparency deficits, limited enforcement capacity, tensions between technical fairness metrics and legal discrimination standards, and the absence of a gender-sensitive and gender-responsive approach across much of the EU digital acquis.

From discriminatory recruitment tools and biased predictive systems to opaque automated decision-making – many examples of which have been set out in the previous chapters of this study – the risks posed by AI to non-discrimination and equality have become a defining concern of the EU's digital

agenda. This chapter maps the current European framework for protection against gender discrimination and TFGBV by AI systems. It also examines how the non-discrimination acquis interacts with the EU's evolving digital regulatory framework, highlighting areas of overlap, complementarity and tension.

Section 5.1 examines the scope of EU gender equality and anti-discrimination law, the prevailing forms of algorithmic discrimination, and the conceptual difficulties arising where algorithmic gender discrimination transcends traditional legal categories. Section 5.2 provides a non-exhaustive overview of the relevant EU digital acquis, analysing the risk-based architecture of the Artificial Intelligence Act alongside the GDPR, the DSA, and Directive (EU) 2024/1385 on combatting violence against women and domestic violence, with a particular focus on their relevance to gender bias and discrimination. Section 5.3 identifies key implementation challenges.

## 5.1. The anti-discrimination framework

### 5.1.1. The scope of EU gender equality and anti-discrimination law

At primary law level, Article 2 of the Treaty on European Union (TEU) lists equality between women and men among the founding values of the EU. Article 3(3) of the TEU identifies the promotion of equality between women and men as one of the EU's missions. Article 8 of the Treaty on the Functioning of the European Union (TFEU) compels the EU to promote the equality between men and women in all EU activities. The TFEU prohibits discrimination on the grounds of sex in Article 19, among the other protected grounds of discrimination of EU law – racial or ethnic origin, religion or belief, disability, age and sexual orientation. Article 20 of the Charter of Fundamental Rights of the European Union confirms that everyone is equal before the law, while Article 21 of same Charter prohibits discrimination on an open list of grounds, including all the grounds mentioned above. This prohibition of discrimination in the Charter is free-standing in formulation but limited in its application as it can only be invoked in matters falling within the scope of EU law. EU law also reflects, at least in part, the substantive understanding of equality adopted in this study (Section 2.3). Article 23 of the Charter requires equality between women and men to be ensured in all areas and allows measures providing for specific advantages in favour of the under-represented sex, while Article 157(4) TFEU permits such measures 'with a view to ensuring full equality in practice' in working life. The concept of indirect discrimination, discussed in Section 5.1.2, similarly looks beyond identical treatment to its effects on particular groups. Nonetheless, the Court of Justice has interpreted positive action narrowly (C-450/93 *Kalanke*; C-409/95 *Marschall*), so that EU law continues to move between formal and substantive conceptions of equality.

At secondary law level, protection against discrimination is provided through a set of directives that have varying scopes of application. The Racial Equality Directive offers the broadest protection, prohibiting discrimination on the grounds of race and ethnic origin in employment, social protection, education, and access to and supply of goods and services (EU 2000/43/EC, 2000). Protection against gender discrimination is primarily provided through the recast Gender Equality Directive (EU 2006/54, 2006) in the field of employment and occupation, and through the Gender Goods and Services Directive (EC, 2012) in relation to access to and supply of goods and services, excluding media, advertising and education. The Employment Equality Directive (EU 2000/78/EG, 2000) establishes a general framework for combatting discrimination on the grounds of religion or belief, disability, age, and sexual orientation, but applies exclusively in regard to employment and occupation. The result of these different directives all addressing non-discrimination separately is that EU non-discrimination law provides an asymmetrical level of protection across protected grounds at EU level, meaning that protection is stronger in certain domains than in others. This asymmetry is particularly relevant for

intersectional discrimination, as disparities in the material scope of EU anti-discrimination directives mean that some combinations of protected grounds, such as gender and disability or gender and religion, are not protected to the same extent as gender and racial or ethnic origin.

Table 2: Scope of the EU Equality directives

| | Sex | (So-called) Race/Ethnicity | Sexual orientation | Religion beliefs | Age | Disability |
|---|---|---|---|---|---|---|
| Education | | ✓ | | | | |
| Goods and services | ✓ | ✓ | | | | |
| Social protection | ✓ | ✓ | | | | |
| Employment | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Source: Based on Xenidis, 2017.

Article 53 of the Charter confirms that Member States may provide a higher level of protection against discrimination under their national laws. Many Member States do so, extending protection beyond EU minimum standards in both the material scope and the personal scope of application – that is, protecting against discrimination on more grounds and in more areas than the EU legal framework requires (EC, 2025d). A Horizontal Equal Treatment Directive, first proposed in 2008 to extend protection beyond employment, has been pending in the Council for seventeen years. Although the EC announced its withdrawal in February 2025 (EC, 2025c), it reversed course in July 2025 and placed the Directive back on the legislative agenda (EP, 2025). The Directive continues to await the unanimous Council support required under Article 19 TFEU to be adopted into the EU acquis.

In addition to providing substantive rules on combatting discrimination, the Equality Directives require Member States to mandate national bodies to promote and protect equality at their national level. These are the so-called Equality Bodies. This institutional enforcement framework has recently been strengthened with the adoption of two Directives (EU) 2024/1499 and 2024/1500 seeking to reinforce the role and capacity of Equality Bodies across Europe (EC, 2025d).[2]

Even where the personal or material scope of EU protection is limited, Member States remain subject to broader obligations under international human rights law.

### 5.1.2. Prevailing forms of algorithmic discrimination

While anti-discrimination law distinguishes between various legally recognised forms of discrimination, including direct and indirect discrimination, harassment and instruction to discriminate, algorithmic gender discrimination often transcends these categories. This section therefore introduces the forms of discrimination most commonly identified in the literature and practice.

[2] Member States had until 19 June 2026 to transpose the Directives into their national law. Note that the EC has opened infringement proceedings against 22 Member States – Belgium, Czechia, Denmark, Germany, Estonia, Ireland, Greece, Spain, France, Croatia, Cyprus, Lithuania, Luxembourg, Hungary, Malta, the Netherlands, Austria, Poland, Portugal, Romania, Slovakia and Finland – for failure to communicate national transposition measures.

### a. Direct versus indirect discrimination

Direct discrimination occurs when 'one person is treated less favourably than another is, has been or would be treated in a comparable situation on grounds of' a protected characteristic. A problem with this conceptualisation of equality is that it is unable to redress more complex forms of structural inequality or proxy discrimination (Gerards and Xenidis, 2021). Indirect discrimination, which is more suitable to address both structural inequality and proxy discrimination, 'occurs where an apparently neutral provision, criterion or practice would put persons of a protected group at a particular disadvantage compared with other persons, unless that provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary'. Direct and indirect discrimination are subject to distinct justification regimes, with indirect discrimination being assessed under the open-ended objective justification test, which affords a broader scope for justification. There is broad consensus that the concept of direct discrimination does not readily accommodate many forms of algorithmic discrimination. The nature of algorithmic discrimination blurs the distinction between direct and indirect discrimination which may reduce legal protection and certainty because direct discrimination and indirect discrimination have distinct justification mechanisms, resulting in the generalisation of the so-called open-ended objective justification test applicable to indirect discrimination (Wachter et al., 2021; Gerards and Xenidis, 2021). This dilutes the more limited grounds for justification that are traditionally available in cases of direct discrimination.

Gerards and Xenidis further contend that the concept of direct discrimination is particularly important in the context of algorithmic discrimination, as it enables reliance on the forms of protection afforded by discrimination by association, as recognised by the Court of Justice of the European Union (CJEU) in Coleman (CJEU, 2008), as well as discrimination by perception, ascription or assumption (Gerards and Xenidis, 2021).[3] These extensions of direct discrimination are especially relevant in cases of algorithmic proxy discrimination and algorithmic misclassification. Algorithmic systems commonly infer characteristics or group memberships from behavioural, demographic or contextual data and subsequently act upon those inferences. Where such predictions are inaccurate, individuals may experience adverse treatment on the basis of characteristics that are merely attributed to them, rather than actually possessed, thereby attracting the protection afforded by the concepts of discrimination by association, perception or assumption.

### b. Intersectional discrimination

Given the data-driven and predictive nature of AI systems, which identify patterns and make inferences from large volumes of data, including proxy variables, these systems may infer multiple protected characteristics simultaneously and classify individuals accordingly. As a result, algorithmic discrimination is likely to be intersectional in nature, that is to involve a combination of several protected grounds (see also Chapter 2 for more background on intersectional discrimination). Intersectional discrimination is one form of multiple discrimination, alongside sequential discrimination, where different grounds operate on separate occasions, and additive (or cumulative) discrimination, where discrimination on multiple grounds occurs in the same situation but each ground remains independently identifiable. Intersectional discrimination, by contrast, arises from the interaction of protected grounds and results in a distinct form of disadvantage that cannot be reduced to, or

[3] The European Commission has promoted such an interpretation by expressing the opinion that the Racial Equality Directive and the Equal Treatment Directive 'also prohibit a situation where a person is directly discriminated against on the basis of a wrong perception or assumption of protected characteristics'. The Council of Europe defines discrimination in CM/Rec(2026)1 on equality and AI: 'Discrimination can be based on a variety of actual or perceived personal characteristics or status, such as [...]'.

established on the basis of, any ground considered in isolation (Fredman, 2016). The EU has increasingly acknowledged the relevance of intersectional discrimination, albeit primarily in recitals and policy documents, rather than in legislation. The Directive on combatting violence against women and domestic violence (see Section 5.2.3) refers to intersectional discrimination in connection with Article 21 of the Charter in multiple provisions (Articles 16, 21, 33) and recognises that persons affected by intersectional discrimination are at a heightened risk of experiencing gender-based violence (Recital 6). Secondly, Article 3(2)(e) of the Directive 2023/970, commonly known as the Pay transparency Directive, provides the first definition of intersectional discrimination in EU law. It covers only combinations that include sex, and only in the field of pay. A third instrument is Directive (EU) 2024/1500 on standards for Equality Bodies, which expressly refers to intersectional discrimination in Article 5 and Recital 5. At the national level, explicit legal recognition of intersectional discrimination remains rare. To date, only Belgium (since 2023) and Spain (since 2022) expressly prohibit it in their legislative frameworks (Chopin and Germaine, 2026). Union law does not preclude the CJEU from considering intersectional discrimination, but the Court has not yet fully addressed or developed its approach to this issue (Weerts et al., 2023). This limited legal recognition is a significant constraint in addressing algorithmic discrimination, which manifests through the interaction of multiple protected grounds.

### c. Proxy discrimination

In this context, 'proxies' refer to characteristics, features or combinations thereof that are not themselves listed among the explicitly prohibited grounds of discrimination under the relevant legal framework, but that correlate with those grounds in a way that their use can lead to unlawful direct or indirect discrimination. The open-ended list of protected grounds in Article 21 of the Charter, together with broader national anti-discrimination laws, assumes particular importance, as it may capture forms of algorithmic discrimination that fall outside the grounds expressly protected under EU secondary anti-discrimination legislation. The qualification of direct or indirect discrimination depends on how inextricably linked the proxy is with a protected ground (CJEU, Dekker – 177/88, 1990; Weerts et al., 2023). Recommendation CM/Rec (2026)1 of the Committee of Ministers to Member States on equality and AI of the COE specifies that 'discrimination can occur by proxy, which occurs on the basis of a seemingly neutral piece of information that is nevertheless related to or correlated with a protected characteristic'. Currently, proxy discrimination is not recognised in legislation at EU level, nor meaningfully implemented in legislation at Member State level. It has been successfully addressed in Courts under the qualification of direct or indirect discrimination. Proxies for gender could include height, pregnancy, sex-based retirement age, literacy, income, career breaks and language, among other things (COE 2026; Weerts et al. 2024; AlgorithmWatch 2022).

### d. Discrimination on unprotected grounds

It is currently debated in academic literature whether the list of protected grounds in the EU anti-discrimination directives is still up to date and appropriate for addressing harms created by algorithmic profiling based on groups that do not map to or correlate with legally protected groups (Wachter et al., 2021a; Bartoletti and Xenidis, 2023; Weerts et al., 2023). AI systems sort individuals into data-driven categories based on, among other things, behavioural patterns, socio-economic circumstances, family composition, educational trajectories, health indicators or consumption profiles. Subsequently, they use these classifications to inform decisions concerning access to opportunities, services and resources. As a result, algorithmic systems may perpetuate or exacerbate existing inequalities related to characteristics such as socioeconomic status, education, health status, welfare dependency, and income, which is particularly problematic when these characteristics are not protected under (national) non-discrimination laws (Gerards and Xenidis, 2021).

#### e. Instruction to discriminate

Scholars have proposed that 'instruction to discriminate' – which is a form of discrimination recognised in the relevant directives – could be suitable to effectively mitigate the substantive and procedural difficulties that arise in the context of algorithmic discrimination (see also Section 7.2.6) (Gerards and Xenidis, 2021; The Gender Equality and Equal Treatment Commissioner's Office of Estonia, 2024).

#### f. Gender-based violence as a form of discrimination against women

Gender-based violence has been conceptualised as a form of discrimination against women. At the EU level, the preamble to Directive on combatting violence against women describes violence against women as 'a persisting manifestation of structural discrimination against women, resulting from historically unequal power relations' between women and men. This reflects an understanding established at the international and the European level. As early as 1992, the Committee on the Elimination of Discrimination against Women (CEDAW) Committee recognised that the definition of discrimination includes gender-based violence, "that is, violence that is directed against a woman because she is a woman or that affects women disproportionately" (CEDAW Committee, 1992, para. 6), a position it reaffirmed in 2017 (CEDAW Committee, 2017). Within the Council of Europe, the Istanbul Convention defines violence against women as "a violation of human rights and a form of discrimination against women" (Article 3(a)), and the European Court of Human Rights (ECtHR) has held that gender-based violence can amount to discrimination under Article 14 of the European Convention on Human Rights (Opuz v. Turkey, 2009). This legal understanding mirrors the conceptual link between gender bias, TFGBV, and substantive equality set out in Section 2.3.

## 5.2. The European Union's digital acquis

For a long time, the legal framework concerning non-discrimination and data protection laws was the primary legal foundation for addressing instances of algorithmic discrimination within the European context. However, recent legal, institutional, and political developments have significantly reshaped the governance landscape of AI in Europe. This section provides an overview of the evolving European regulatory framework relevant to algorithmic discrimination, especially regarding AI. Given its significance as the first comprehensive EU legal framework governing AI, particular attention is devoted to the AI Act and its implications for gender equality and non-discrimination. The section further briefly examines the GDPR, the DSA, and the Directive on combatting violence against women, as well as relevant legal instruments introduced by the COE. It concludes by outlining forthcoming legislative developments that may have an impact on the protection against algorithmic discrimination and gender-based harms in the future.

### 5.2.1. The Artificial Intelligence Act

In the spring of 2024, the EU formally adopted Regulation (EU) 2024/1689 on harmonised rules on artificial intelligence, commonly known as the AI Act. The AI Act forms part of the EU's New Legislative Framework for product regulation and is stated to create a comprehensive and technology-neutral EU legal regime across sectors. The AI Act is a so-called risk-based regulatory framework. This means that the Act requires providers of AI systems to assess the level of risks their systems pose to health, safety, and fundamental rights before placing them on the market. Systems are classified as unacceptable

risk,4 high risk5 and limited or minimal risk. Each classification carries different legal requirements, with most of the obligations set out in the AI Act applying to high-risk systems. General-purpose AI (GPAI) models are subject to a specific legal regime and distinguished based on the systemic and non-systemic nature of the risks they pose (Articles 3(63), Articles 21, 53 and 55 of the AI Act). In addition, when GPAI is integrated within an AI system, that system is also subject to the legal requirements applicable to its risk category. Specific provisions applicable to high-risk AI systems addressing algorithmic discrimination are discussed in Chapter 6.

While the AI Act does not contain a standalone provision specifically dedicated to non-discrimination, the risk-based regulatory framework is premised on the assessment that AI systems may pose risks to health, safety, and fundamental rights. The latter explicitly encompasses the right to non-discrimination, which includes discrimination on the basis of gender. The AI Act acknowledges that the notion of 'risk' is not limited to technological characteristics. It is tied to the societal context in which AI systems operate. In this regard, the recitals recognise that certain uses of AI systems can lead to discrimination between persons or groups and may perpetuate historical patterns of discrimination, such as those based on racial or ethnic origins, gender, disabilities, age or sexual orientation, or may create new forms of discriminatory impacts (Recitals 56, 57, 58 and 67). Beyond the references to non-discrimination, the AI Act contains several references to gender equality and the risk of historical patterns of discrimination against women is recognised in specific domains and applications.6 This has prompted the question of why gender equality is not recognised as a transversal governance consideration (Karagianni, 2025), given that algorithmic gender bias is not limited to the contexts expressly recognised in the Regulation. The AI Act does not refer to transgender, non-binary or intersex and gender non-conforming people specifically, nor does it specifically address the risks and impacts that AI systems may pose for these groups. Some scholars have therefore observed that the AI Act does not comprehensively address inclusive gender identities, perpetuating a binary distinction of gender (Karagianni, 2025). The understanding of non-discrimination as a fundamental rights risk has implications throughout the Regulation. It not only informs the substantive obligations attached to each risk category but is also reflected in the broader governance and enforcement framework that the Regulation establishes. An example is the distinctive cooperation mechanism established between national market surveillance authorities (MSAs) and national public authorities or bodies that supervise or enforce adherence to the obligations under Union law that protect fundamental rights (Article 77). The explicit inclusion of the right to non-discrimination in Article 77 of the AI Act embeds equality considerations within the enforcement architecture and the cooperation framework between these authorities (see also 7.2.4).

The list of prohibited AI practices and the list of high-risk AI systems in Annex III of the AI Act reflect the EU legislature's assessment of technological capabilities, societal concerns and regulatory priorities at the time of the AI Act's adoption. As such, the lists are limited to the risks identified by the EU legislature at that time. Recognising this, the Regulation subjects both lists to review procedures.7 The

---

4 Recital 28 explains why AI systems can be prohibited under the AI Act: 'they contradict Union values of respect for human dignity, freedom, equality, democracy and the rule of law and fundamental rights enshrined in the Charter, *including the right to non-discrimination,* to data protection and to privacy and the rights of the child' [emphasis added].

5 Recital 31 explains why systems are considered high risk under the AI Act: 'they may violate the right to dignity and *non-discrimination* and the values of equality and justice' [emphasis added]; Recital 48 further specifies: '[t]he extent of the adverse impact caused by the AI system on the fundamental rights protected by the Charter is of particular relevance when classifying an AI system as high risk'.

6 The AI Act contains explicit references to gender equality in Recitals 27 and 48 and Article 95(2)(e). Historical patterns of discrimination against women are recognised in Recitals 56 and 57, whereas historical patterns of discrimination based on gender are recognised in Recital 58.

7 The EC can adopt delegated acts to amend the list of high-risk use cases. Article 7(1) first sub-paragraph: add or modify use-cases of high-risk AI systems listed in Annex III Article 7(3) first sub-paragraph: remove high-risk AI systems listed in Annex III. High-risk AI systems cannot be added. Prohibited practices can be updated via legislative amendment. The EC assesses 'the need for amendment of

criteria that the EC must take into account before adding or modifying use-cases, and removing high-risk AI systems from Annex III, indicate that fundamental rights are a central consideration in any amendment to Annex III, including non-discrimination and including on the basis of gender (Article 7(2)).

Experts have cautioned that, in relation to both the prohibition practices and the high-risk regime, there is a risk of so-called 'de-risking practices' (Shrishak and Pénicaud, 2025). These practices occur when providers or deployers of AI systems conclude that one of the cumulative conditions for the application of a prohibition is not fulfilled, or when they classify systems to be outside the high-risk regime. This can happen even if these systems pose significant risks to fundamental rights, among other things. Such practices may result in AI systems falling outside the scope of the AI Act, thereby limiting the applicability of the safeguards and protections established in it. This would undermine the effectiveness of the AI Act by enabling certain systems to circumvent the regulatory obligations and protections that it seeks to ensure. The AI Act applies horizontally without prejudice to other Union legislation (Article 7(2)). Accordingly, AI practices that do not qualify as prohibited practices under the AI Act may nevertheless be prohibited under other provisions of Union law (Article 5(8)). This means that compliance with the AI Act is not a sufficient condition for compliance with non-discrimination law or data protection law and an AI system not prohibited under the AI Act may nonetheless be unlawful under other primary or secondary EU law (EC, 2025b). Consequently, compliance with EU non-discrimination law is required irrespective of whether an AI system is classified as prohibited, high-risk, limited-risk or falls outside the scope of the AI Act's risk-based categories (Recital 27 AI Act).

Box 1: Concretisation of risk classification in the AI Act

**Article 5 of the AI Act prohibits nine practices**:

1. harmful AI-based manipulation and deception,
2. harmful AI-based exploitation of vulnerabilities,
3. social scoring,
4. individual criminal offence risk assessment or prediction,
5. untargeted scraping of internet or CCTV material to create or expand facial recognition databases,
6. emotion recognition in workplaces and education institutions,
7. biometric categorisation to deduce certain protected characteristics,
8. real-time remote biometric identification for law enforcement purposes in publicly accessible spaces,
9. AI systems that generate non-consensual sexually explicit and intimate content, such as 'AI nudification apps' or child sexual abuse material (CSAM).

Prohibitions 1-8 came into effect in February 2025. Prohibition 9 comes into effect in December 2026 and was introduced as a part of the AI Omnibus.

Prohibitions address risks of algorithmic discrimination through bans of certain AI systems; prohibitions only apply when several cumulative conditions are met.

the list of prohibited AI practices' once a year (Article 74(2)) and submits the findings to the European Parliament and the Council (Article 112(1)).

**Annex III lists High-Risk AI Systems referred to in Article 6(2) of the AI Act**:

1. biometrics,
2. AI safety components in critical infrastructures (digital infrastructure, road traffic, supply of water, gas, electricity),
3. education and vocational training,
4. employment, workers' management and access to self-employment,
5. access to and enjoyment of essential private services and essential public services and benefits (e.g. social security, credit scoring, insurance, healthcare, emergency services),
6. AI use-cases in law enforcement,
7. AI use-cases in migration, asylum and border control management,
8. AI solutions used in the administration of justice and democratic processes.

Starting on 2 December 2027, systems referred to in Annex III will be considered high-risk AI systems and will be subject to the requirements for such systems laid down in the AI Act (Article 6 (2) AI Act). The AI Act provides for a derogation: providers of AI systems in these areas listed in Annex III can opt out of being considered high-risk if they consider that their specific case does not pose a 'significant risk'. In that scenario, the obligations for high-risk systems do not apply (Article 6 (3) AI Act). Systems profiling natural persons are always considered to be high-risk systems (Article 6 (3) AI Act).

Source: EU Artificial Intelligence Act (Regulation (EU) 2024/1689) and related implementation, guidance and Digital Omnibus on AI materials.

Of particular relevance to this study is the introduction, through the Digital Omnibus on AI, of a prohibited AI practice covering AI systems that generate non-consensual sexually explicit or intimate content. The mass, non-consensual generation of sexual imagery is a well-documented phenomenon, evidenced by cases involving public figures such as Taylor Swift, incidents targeting minors, and the proliferation of "nudification" applications and websites, culminating in the so-called "Grok scandal" of 2026, involving the large-scale generation of sexually explicit images depicting identifiable women and girls through Grok, xAI's generative AI system integrated into X (AI Forensics, 2026). Legal frameworks have generally struggled to keep pace with this abuse: only a small number of jurisdictions prohibit the creation of such content outright, leaving many of these technologies to operate in a legal grey area.

The prohibition on AI systems that generate or manipulate non-consensual intimate imagery, Article 5(1)(ba) AI Act, introduced through the Digital Omnibus on AI and applicable from 2 December 2026, is a welcome addition. It closes a gap left by Article 50, whose transparency obligations protect the viewers of synthetic content but not the individuals depicted within it, and responds to the disproportionate targeting of women and girls by this technology (Recital 11) (Rigotti, 2026b). The prohibition is not without limitations, however. Substantively, it covers only AI systems that create realistic depictions of an identifiable person's intimate parts or of a person engaged in sexually explicit activity, with "intimate parts" defined in strictly anatomical terms (Recital 12). As a result, the production and manipulation of sexualised imagery such as a woman depicted in a bikini falls outside its scope. Moreover, this stereotypically Western, anatomical conception of intimacy is not culturally sensitive: removing religious attire such as a niqab causes comparable harm to its victims, yet risks escaping the prohibition unless "intimate parts" is read expansively. Procedurally, the AI Act regulates providers and deployers of AI systems, not end-users as such: an individual nudifying a photo is only

caught insofar as they qualify as a "deployer", and the "household exception" of Article 2(10) excludes AI systems used in a personal, non-professional capacity, although, sharing the content, or profiting from it, may bring the creator back within scope as a deployer (Rigotti, 2026b). Moreover, providers of intermediary services through which the resulting content circulates benefit from far-reaching legal immunity for material published by their users (see C7). End-user conduct is thus left to other bodies of law, such as criminal law under VAW Directive.

A parallel prohibition, Article 5(1)(bb), extends a similar approach to AI-generated and manipulated child sexual abuse material, also referred to as CSAM.

### 5.2.2. The General Data Protection Regulation

The GDPR (Regulation (EU) 2016/679) is the EU legislation aimed at protecting the personal data of natural persons. The GDPR applies to AI systems and ADM systems that process personal data. For AI systems, the GDPR and AI Act may apply adjacent to each other. The GDPR is particularly relevant in the context of this study given that the GDPR also applies to systems not qualifying as an AI system under the AI Act, and to which the framework of the AI Act set out in Section 5.2.1 does not therefore apply. The GDPR confers important rights in the context of automated decision-making, including rights to information and transparency (Article 15), as well as the right not to be subject to certain solely automated decisions (Article 22). These provisions are discussed in Chapter 7.

Article 9 of the GDPR prohibits the processing of special categories of personal data, including data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data for identification purposes, health data, and data concerning sex life or sexual orientation. As a general rule, this provision prevents providers and deployers of AI systems from using such data as variables (e.g. labels, input variables, risk factors). Article 4(a) of the AI Act allows the processing of such data when the aim is to detect and mitigate algorithmic discrimination. While Article 9 of the GDPR protects certain sensitive categories of personal data, regardless of whether discrimination occurs, the non-discrimination framework protects against discriminatory treatment based on protected grounds, regardless of whether data concerning those characteristics is processed. The distinction is particularly relevant in relation to gender. Although gender is a protected ground under EU equality law and Article 21 of the Charter, it is not included among the special categories of data listed in Article 9(1) of the GDPR. As a result, the safeguards for sensitive data processing laid out in both the GDPR and AI Act do not apply for gender data. This does not mean that gender-related data fall entirely outside the GDPR's protective framework. Some data protection authorities, including in Spain and Portugal, have recognised women as potentially vulnerable data subjects in specific contexts, such as consumer protection and cases involving victims of sex-related crimes (Malgieri and Fuster, 2022). Unlike Article 9(1) GDPR, which attaches enhanced protection to certain categories of data, vulnerability is a contextual and situational notion that attaches to the data subject. It does not prohibit processing but may inform the assessment of risks to data subjects and the safeguards required in a particular context, subject to the assessment of supervisory authorities. While this may mitigate some practical consequences of the exclusion of gender from Article 9(1), it cannot close the protection gap.

The disconnect between the recognition of gender as a protected ground under EU anti-discrimination laws and its omission from the special categories of personal data protected under Article 9(1) GDPR has implications extending beyond data processing. For example, in its draft guidelines on the classification of high-risk AI systems, the EC interprets Point 1(b) of Annex III, concerning biometric categorisation, as referring exclusively to characteristics listed in Article 9(1) GDPR, relying on Recital 54 of the AI Act (EC, 2026). Under this interpretation, biometric categorisation on the basis of

characteristics such as gender – but also age, language, or membership of a national minority – would fall outside the high-risk regime.[8] This raises questions over whether the exclusion of gender and other characteristics from outside Article 9(1) of the GDPR by the European legislator – even though these constitute protected grounds under EU fundamental rights and equality law – is in accordance with the EU fundamental rights framework (EDPS and EDPB, 2021; Malgieri and Fuster 2022).

### 5.2.3. The Digital Services Act

In 2022, the EU adopted the DSA (Regulation (EU) 2022/2065), which aims to contribute to the proper functioning of the internal market by establishing harmonised rules for safe, predictable, and trusted cyberspace. It rests on two pillars, i.e. liability exemptions and due diligence obligations, and applies to providers of intermediary services, such as social media. Respect for gender equality and non-discrimination by these providers forms part of the DSA architecture.[9] The DSA contains provisions on gender discrimination and is therefore relevant for this study. Articles 34 and 35 of the DSA explicitly oblige Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) to identify, assess and mitigate systemic risks to fundamental rights including to non-discrimination, as enshrined in Article 21 of the Charter (Article 34 (b)), and in relation to gender-based violence (Article 34(d)) that stem from the design or functioning of their service and its related systems, including algorithmic systems, or from the *use* made of their services. This includes, for example, algorithmic recommender systems or integrated AI-chatbots. The inclusion of Article 21 of the Charter as a benchmark for systemic risk assessments implies a horizontal and expansive reading of the right to non-discrimination. The precise implications and limits of this approach, however, remain unclear. Part of these assessments is evaluating how VMOPs and VLOSEs may facilitate the dissemination of illegal content, including hate speech. The platforms are required to adopt corresponding mitigation measures.

One limitation of the DSA's risk-management framework is its focus on mitigation of identified harms rather than the prevention of discriminatory outcomes. This reactive logic is ill-suited to addressing structural forms of discrimination embedded in datasets, design choices and optimisation processes, which require preventive and equality-by-design interventions beyond the mitigation measures envisaged in Article 35 DSA. A further challenge relates to the DSA's notion of systemic risk. The DSA targets risks capable of affecting a significant part of the EU population (Recital 86; *Amazon v Commission*, C-367/23), whereas EU equality law is centred on the protection of individual rights irrespective of the scale of the harm concerned. The meaning of a 'significant' impact remains uncertain and, if construed restrictively, may exclude discriminatory practices affecting relatively small or marginalised groups, despite such harms falling squarely within the ambit of anti-discrimination law. Relevant obligations and mechanisms of the DSA are further discussed in Chapter 7.

### 5.2.4. The Directive on combatting violence against women and domestic violence

In May 2024, the EU adopted Directive (EU) 2024/1385 on combatting violence against women and domestic violence (VAW Directive), the first binding EU instrument aimed at harmonising legislative

[8] Recital 16 of the AI Act adopts a considerably broader understanding of biometric categorisation that explicitly encompasses categories beyond Article 9(1) GDPR.

[9] Relevant DSA provisions address both (systemic) discriminatory impacts (including *illegal* hate speech) and the non-arbitrary, non-discriminatory implementation of platform obligations, including content moderation and notice-and-action procedures (Recitals 3, 12, 26, 40, 47, 52, 58, 68, 69, 81, 94, 134; Articles 14(4), 16, 20(4), 26, 34(1)(b), 35 and 48(4)(e)).

and policy responses to such violence across Member States. The deadline for transposition is set in June 2027.

Among its most significant innovations is the criminalisation of several forms of TFGBV: the non-consensual sharing of intimate or manipulated material (Article 5), cyberstalking (Article 6), cyberharassment (Article 7), and cyberincitement to violence or hatred (Article 8). In addition to these offences, the Directive builds on the existing EU framework on victims' rights to strengthen support and assistance for survivors, addresses the specific needs arising from gender-based violence, and places considerable emphasis on prevention and research, while seeking to facilitate cooperation among Member States and contribute to an area of freedom, security and justice (Recital 1). Specifically, the Directive emphasises the documentation of evidence of technology-facilitated abuse (Articles 14(1) and 25(1)), the expertise and investigative tools available to law enforcement (Article 15(1)), the training of professionals working with victims to ensure a specialised and sensitive response (Article 36(10)), the development of digital literacy skills and critical engagement with the online world (Article 34(8)), and regular consultation with non-governmental organisations across victim support, policymaking, awareness-raising, research, training and monitoring (Article 41). Recognising intersectionality, the Directive extends protection and access to improved support services to those most at risk of gender-based violence while being the least supported in society, including racialised women, women with disabilities, transgender and non-binary individuals, sex workers, individuals with lower socio-economic status, homeless individuals, and those with unstable immigration status (Articles 16(4), 21(g), 33(1)).

Nonetheless, the Directive falls short of integrating a clear gender perspective. It omits any definition of 'gender' and fails to draw a conceptual distinction between 'gender' and 'sex'. Instead, the terms are employed interchangeably, leaving the Directive's treatment of gender conceptually indeterminate. This omission matters because the EU acceded to the Istanbul Convention in October 2023, in respect of matters falling within its competence. This means that the Istanbul Convention currently forms an integral part of EU law, ranks above secondary legislation, and the Directive must be interpreted in conformity with it (d'Ursel, 2023). The Istanbul Convention defines gender in Article 3(c) as 'socially constructed roles, behaviours, activities and attributes that a given society considers appropriate for women and men'. The Directive's failure to engage with this established conceptual framework thus represents a notable departure from, rather than an alignment with, the European standards it purports to build upon (McGlynn, 2024).

### 5.2.5. Recent EU legislative and regulatory developments

Several legislative developments warrant attention as they may be transformative for gender equality and non-discrimination protection in Europe. One is the EC's 2022 proposal for an AI Liability Directive (AILD) (EC, 2022). The AILD sought to introduce a rebuttable presumption of causation to ease the burden of proof for claimants seeking compensation for AI-caused harm (see also Chapter 7). The AILD was formally withdrawn from the legislative process in February 2025.

The upcoming Digital Fairness Act (EC, 2025c) aims to address commercial practices enabled by algorithmic systems, including targeted advertising and personalised pricing. One of its stated aims is to have implications for fundamental rights and equality.

In the domain of employment, Directive (EU) 2024/2831 on improving working conditions in platform work, which must be transposed by Member States by 2 December 2026, introduces the first binding EU rules on algorithmic management applicable to digital labour platforms and platform workers, requiring human oversight of automated decisions and transparency towards platform workers

(Council of the European Union, 2026). Additionally, the EP has called on the EC to legislate on algorithmic management at work more broadly, demanding human oversight and transparency for automated decisions affecting workers' terms of employment (EP, 2025).

The EC introduced the Digital Omnibus on AI and ongoing Omnibus on Data Protection in November 2025 as part of its agenda to simplify and consolidate the digital rulebook. The Digital Omnibus on AI Regulation was approved by the Council on 29 June 2026 and postpones the application of obligations for high-risk AI systems until 2 December 2027 or later, depending on the availability of compliance-support tools such as harmonised standards. The Data Omnibus is ongoing at the time of writing. One of the proposed amendments could have far-reaching implications for AI and gender discrimination concerns. The proposed amendment to Article 3(1) would weaken the GDPR's definition of personal data by introducing a relative and subjective approach under which information could be treated as non-personal if a specific entity cannot identify an individual, even where others could or where a future recipient may be able to do so. This, in turn, would narrow the scope of GDPR protections and oversight. Concerns have been raised both by institutional actors and CSOs regarding the potential effects of certain simplifications on transparency safeguards and fundamental rights (NOYB 2026; Equinet and the European Network of National Human Rights Institutions (ENNHRI), 2026; EDPB and EDPS, 2026).

### Box 2: Prohibition on AI-generated non-consensual intimate imagery or the so-called nudifier ban

2026 opened with the so-called "Grok scandal", but the mass, non-consensual generation of sexual imagery was already a well-documented phenomenon, evidenced by cases involving public figures such as Taylor Swift, incidents targeting minors, and the proliferation of "nudification" applications and websites (AI Forensics, 2026). Legal frameworks have generally struggled to keep pace with this abuse: only a small number of jurisdictions prohibit the creation of such content outright, leaving many of these technologies to operate in a legal grey area. Against this background, the prohibition on AI systems that generate or manipulate non-consensual intimate imagery, Article 5(1)(ba) AI Act, introduced through the Digital Omnibus on AI and applicable from 2 December 2026, is a welcome addition. It closes a gap left by Article 50, whose transparency obligations protect the viewers of synthetic content but not the individuals depicted within it, and responds to the disproportionate targeting of women and girls by this technology (Recital 11) (Rigotti, 2026).

The prohibition is not without limitations, however. Substantively, it covers only realistic depictions of an identifiable person's intimate parts or of a person engaged in sexually explicit activity, with "intimate parts" defined in strictly anatomical terms (Recital 12). As a result, sexualised imagery such as a woman depicted in a bikini falls outside its scope. Moreover, this stereotypically Western, anatomical conception of intimacy is not culturally sensitive: removing religious attire such as a niqab causes comparable harm to its victims, yet risks escaping the prohibition unless "intimate parts" is read expansively (Rigotti et al., 2024).

Procedurally, the AI Act regulates providers and deployers of AI systems, not end-users as such: an individual nudifying a photo is only caught insofar as they qualify as a "deployer", and the "household exception" of Article 2(10) excludes AI systems used in a personal, non-professional capacity, although, sharing the content, or profiting from it, may bring the creator back within scope as a deployer (Rigotti, 2026). Moreover, providers of intermediary services through which the resulting content circulates benefit from far-reaching legal immunity for material published by their users (see C7). End-user conduct is thus left to other bodies of law, such as criminal law under VAW Directive.

A parallel prohibition, Article 5(1)(bb), extends a similar approach to AI-generated and manipulated child sexual abuse material, also referred to as CSAM.

### 5.2.6. The Council of Europe's Framework

The COE has introduced several instruments addressing algorithmic discrimination. Most relevant in the context of this study is the Framework Convention on AI, which was adopted in 2024 and has been ratified by the EU on 15 May 2026 (Council of the European Union, 2026). The Framework Convention aims to ensure that activities within the lifecycle of AI systems are fully consistent with human rights, democracy and the rule of law. The Convention adopts a more abstract, principles-based approach compared with the detailed provisions of the AI Act. For the EU, the Convention will be implemented exclusively through the AI Act and other relevant Union acquis, where applicable (Council of the European Union, 2026), thereby linking the international Convention's framework to the Union's own regulatory architecture. Notwithstanding its implementation through the AI Act and other relevant Union acquis, the Convention retains autonomous legal relevance within the EU legal order. As a binding international agreement concluded by the Union, it forms an integral part of EU law and informs the interpretation of secondary legislation, including the AI Act. A particular added value of the Convention lies in its potential international legal reach. The Convention has already attracted the signatures of several non-EU states, including the United States, Canada, Japan, Israel, the United Kingdom, Norway, Iceland and Switzerland. Upon ratification, signatory states would be required to implement the Convention's obligations, including with respect to private-sector activities. At the same time, the Convention permits a graduated and differentiated regulatory approach, allowing obligations imposed on private actors to differ in scope and intensity from those applicable to public authorities (Article 3 (b)).

## 5.3. Implementation challenges

Despite the existence of multiple legal instruments addressing gender bias and discrimination in AI systems, significant implementation challenges remain. The effective implementation of the legal framework discussed above is constrained by several structural and mutually reinforcing barriers. These barriers include limited transparency, enforcement deficits, conceptual misalignments between non-discrimination law and algorithmic fairness, the limited integration of gender-sensitive and gender-responsive considerations across the regulatory framework, and tensions between European attempts at harmonisation and differing national approaches to fundamental rights protection.

First, persistent transparency deficits impede the effective detection and redress of discriminatory outcomes arising from AI systems by limiting the ability of affected individuals, civil society organisations and enforcement bodies to identify potentially discriminatory practices (Smuha, 2021).

Second, the effective enforcement of the relevant regulatory frameworks is constrained by institutional and operational capacity limitations. The cross-border deployment of AI and other digital systems frequently gives rise to complex jurisdictional and coordination challenges, requiring cooperation between multiple national authorities operating within different legal and administrative contexts. These challenges are compounded by the persistent constraints on resources, expertise and capacity that affect many competent authorities (EDPB, 2021). These issues are further examined in Chapter 7.

A third challenge arises from the conceptual mismatch between technical approaches to algorithmic fairness and the legal framework (Wachter et al., 2021a, Gerards and Xenidis, 2021, Weerts et al., 2023). The AI Act does not define discrimination, nor does it provide guidance on how discriminatory outcomes should be identified and assessed in practice. Whereas AI systems typically identify patterns and generate predictions based on correlations rather than causation, EU non-discrimination law requires a contextual assessment of adverse treatment or effects, including the identification of appropriate comparators and the consideration of potential justifications (Wachter et al., 2021 a,b). As a result of the high degree of individualisation and differentiation enabled by AI systems, discriminatory

outcomes may not always be identified by technical fairness metrics, highlighting a persistent gap between technical assessments of fairness and legal assessments of discrimination. This will be discussed further in Chapter 6.

A further issue is that most of the EU digital acquis was adopted without a gender-sensitive and gender-responsive approach. Gender equality is seldom addressed explicitly and, where it is, this is formulated in general terms rather than translated into specific obligations. As a result, much is left to interpretation and application, so that the protection these legal instruments offer against gender bias in AI depends on the choices of those who enforce them. For example, in its enforcement activities concerning the risk assessments that providers must carry out under Articles 34 and 35 DSA, the EC has so far paid little attention to TFGBV (Rigotti et al., 2026). Similarly, the EU response to the creation and dissemination of sexual digital forgeries has been slow and largely reactive: a prohibition was introduced through the Digital Omnibus on AI only after the Grok scandal.

Finally, the effective integration of fundamental rights considerations into the EU digital acquis is complicated by tensions between regulatory harmonisation via the maximum harmonisation principle and national traditions of fundamental rights protection. While Recital 9 of the AI Act explicitly states the applicability of the AI Act without prejudice to existing EU fundamental rights law, and while Article 53 of the Charter preserves the possibility for Member States to maintain more extensive protections, the practical implementation of the framework may nevertheless favour convergence around a common regulatory baseline. This dynamic is particularly visible where key compliance requirements are specified in harmonised standards. The AI Act delegates development of substantive requirements, such as technical detail risk-management systems, accuracy metrics, and human oversight through harmonised standards by the standardisation bodies CEN and CENELEC (see Section 2.3). Compliance with the harmonised standards will give rise to a presumption of conformity under Articles 40 and 41 of the AI Act. In many Member States, national equality legislation extends beyond the protection required under EU equality directives, reflecting a broader personal and material scope of application (Chopin and Germaine, 2026).[10] The result is an asymmetry between formal regulatory autonomy and practical regulatory discretion: Member States remain free, in principle, to maintain and give practical effect to more extensive anti-discrimination laws, yet the central role of harmonised standards in establishing conformity may reduce the practical capacity for Member States to do so.

[10] For an overview of non-discrimination law in EU Member States, see the comparative analysis for by the European network of legal experts in gender equality and non-discrimination, https://doi.org/10.2838/6732664.

# 6. MAPPING OF TECHNICAL, ORGANISATIONAL AND GOVERNANCE MEASURES TO PREVENT AND MITIGATE GENDER BIAS

KEY FINDINGS

- The AI Act prescribes no specific debiasing techniques or fairness metrics.
- Since conformity assessments are mostly self-assessments by providers, effectiveness in practice depends on the compliance and risk management processes they embed internally.
- Bias rarely stems from a single factor but from multiple interacting ones, and technical interventions are inherently limited: removing sensitive variables or neutralising gendered language does not suffice for eliminating bias, as proxy variables continue to encode protected characteristics, and mitigating one form of bias may reinforce another. Available techniques intervene at different points of the AI lifecycle: pre-processing, in-processing and post-processing.
- No fairness metric suits all circumstances: the metric chosen determines which disparities become visible, and the same system may appear fair or unfair accordingly. Whether residual risk is 'acceptable' is left to the provider, which is contested in academic literature.
- At governance level, the FRIA complements providers' obligations, as discriminatory impacts often emerge in the deployment context. Yet its scope is limited to public-sector and specific financial deployers, and participatory methods are not required even though they can identify gender-specific and intersectional harms. Impact assessments also serve a preventive function: they may conclude that a system should not be developed or deployed at all.
- A gap exists between the EU's established gender mainstreaming framework and its digital regulatory framework, where gender impact assessment methodologies remain largely absent.
- AI can advance equality, but this cannot be presumed and, in practice, there are more examples of AI exacerbating existing inequalities than AI advancing equality. Algorithmic discrimination typically results from design choices, data limitations or reproduced societal inequalities rather than deliberate exclusion. Realising AI's benefits requires deliberate, rights-based and gender-responsive regulation and governance.
- Best practices at European and national level largely sit within broader equality and fundamental rights frameworks. Practical assessment tools have been developed at EU level and, among others, in Belgium, Finland, Portugal, Estonia, Lithuania, Denmark and the Netherlands. Several Member States stand out for interdisciplinary and interinstitutional cooperation: Spain integrates gender equality across the AI supervision cycle, France provides technical expertise to its equality body, Denmark centres stakeholder engagement.

This chapter maps the technical, organisational, and governance measures available to prevent and mitigate gender bias in AI systems. This tripartite focus reflects the approach to trustworthy AI advanced by the AI HLEG (HLEG, 2019), which emphasises that bias and discrimination cannot be avoided through technical design choices alone, but require complementary organisational processes and governance mechanisms throughout the AI lifecycle. It also builds on the distinction between pre-existing, technical, and emergent bias drawn in Section 2.4.

Section 6.1 sets out the requirements of the AI Act. Section 6.2 provides an overview of several state-of-the-art technical bias mitigation measures and fairness metrics, illustrated through the application of a case study. These measures primarily address technical bias. Pre-existing and emergent bias, however, cannot be identified or mitigated by technical means alone, thereby requiring organisational and governance instruments, including Fundamental Rights Impact Assessments (FRIAs), which are addressed in Section 6.3. This chapter closes with a selection of best practices concerning technical, organisational and governance measures to prevent and mitigate gender bias at European and national level in Section 6.4.

## 6.1. AI Act requirements

As discussed in Chapter 5, the AI Act adopts a risk-based approach and establishes a range of obligations aimed at identifying and mitigating risks of bias, particularly in relation to high-risk AI systems. The AI Act requires providers to self-assess compliance with legal requirements before products are placed on the market. When assessing the effectiveness and practical significance of the bias mitigation measures discussed in this chapter, it is important to keep in mind that in the majority of cases the provider is thus primarily responsible for conformity assessments. External oversight by notified bodies is only required in limited circumstances. As a consequence, whether the gender bias mitigation measures prescribed by the AI Act and other relevant legislation are effective in practice depends considerably on the measures that are embedded within the AI system's providers' compliance- and risk-management processes.

The AI Act does not prescribe specific bias mitigation techniques (see section 6.2.1.) or fairness metrics (see section 6.2.2.). Article 9 of the AI Act requires providers of high-risk AI systems to establish, implement and maintain a risk management system. This includes the identification, analysis and mitigation of risks to health, safety and fundamental rights. Article 10 of the AI Act, which concerns data and data governance, lays down quality requirements for training, validation, and testing datasets. This is the AI Act's main provision dealing with issues relating to bias and discrimination. Article 10(2)(f) specifically requires an 'examination in view of possible biases that are likely to [...] lead to discrimination [...]'. Providers must take 'appropriate measures to detect, prevent and mitigate possible biases identified according to point (f)' pursuant to Article 10(3). Article 4a of the AI Act, which was introduced by the Digital Omnibus, amends the safeguards governing the processing of sensitive data that were originally set out in Article 10(5). Under Article 11 of the AI Act, providers must draw up technical documentation, with Annex IV specifying what information must be included at a minimum. This includes the intended purpose of the system (Annex IV, para. 2(a)), human oversight measures (Annex IV, para. 2(e)), information on the datasets used, including datasheets where appropriate (Annex IV, para. 2(d)), as well as bias testing, debiasing methods and fairness metrics (Annex IV, para. 2(g)). Providers must also document the expected level of accuracy, including, where relevant, degrees of accuracy for specific persons or groups of persons (Annex IV, para. 3). Similar technical documentation obligations apply to providers of GPAI under Article 53 of the AI Act and Annex XI. In addition, Article 27 of the AI Act requires certain deployers of high-risk AI systems to carry out a FRIA.[11] Finally, Articles 49(1) and (2) of the AI Act require the registration by providers of certain high-risk AI systems in the EU database established under Article 71. Summaries of FRIAs must be submitted as well (Annex VII Section C (4)).

[11] The assessment consists of a description of the deployer's processes that includes: (a) the period of time within which and the frequency with which the AI system will be used; (b) the categories of natural persons and groups likely to be affected by use of the AI system in the specific context; (c) the specific risks of harm likely to impact those groups; (d) human oversight measures taken; and (e) how those risks are addressed, including through internal governance and complaint mechanisms.

## 6.2. Socio-technical measures to prevent and mitigate gender bias

As explained in Chapter 2, the AI Act does not define bias or discrimination. JTC-21, the CEN-CENELEC standardisation committee tasked with developing harmonised standards under the AI Act, considers unwanted bias to be context-dependent: whether a bias is 'unwanted' depends on the specific AI system and the assessment of the AI system's provider or deployer.

Technical interventions are rarely straightforward. As discussed in Chapter 2, merely removing sensitive variables such as direct references to someone's gender from the dataset is not an effective approach to mitigate bias. Among other reasons, this is because proxy variables may continue to encode information about protected characteristics (Dwork et al., 2012; Kamiran and Calders, 2012). Likewise, replacing gender-specific words with gender-neutral words is unlikely to mitigate all bias (Gonen and Goldberg, 2019). The most appropriate approach to combatting gender bias depends on the technical specifics of the AI system, the data provenance, and the context in which the AI system is developed and used. Moreover, reducing one type of bias may unintentionally introduce or reinforce another (Mittelstadt, 2024). Given the socio-technical nature of AI systems, technical approaches are inherently limited and cannot fully address bias (Buyl and De Bie, 2024; Schwartz et al., 2022; Wachter et al., 2021b). To effectively address bias, contextual and socio-cultural knowledge should complement technical approaches. This will be further discussed in section 6.3.

Identifying the sources of bias is the first step in any strategy for mitigating unwanted bias. To assess and mitigate bias, a variety of fairness metrics and bias mitigation measures have been developed. While the AI Act does not prescribe specific fairness metrics or bias mitigation techniques, providers and deployers may use such tools to demonstrate, assess and monitor compliance with the Regulation's risk-management, data-governance and fundamental-rights objectives. Before discussing specific bias mitigation measures, a fictitious example is introduced that will be used throughout the chapter to illustrate how the measures may work in practice.

Box 3: Fictious example used throughout the chapter

An AI-assisted recruitment system screens CVs and ranks candidates for interview invitations. Historical training data reflect past and current gender and racial imbalances in hiring for management roles. The system has learned that career breaks are correlated with poor future job performance. This learned bias could inadvertently disadvantage candidates who have taken maternity or paternity leave. In addition, candidates from racialised backgrounds, particularly women of colour, are under-represented in the training data due to historical hiring practices. As a result, the model has fewer examples from which to learn patterns associated with successful candidates from these groups. The unwanted negative associations between parental status and merit, as well as the underrepresentation of racialised candidates in the training data, are examples of the types of harms that debiasing and fairness methods aim to address.

### 6.2.1. Bias mitigation measures

Once potential bias or unfairness has been identified, various technical interventions can be applied throughout the AI lifecycle to mitigate this bias or unfairness. Bias can be mitigated through interventions at different stages of the AI lifecycle. Accordingly, bias mitigation methods are commonly classified into three categories: pre-processing techniques, in-processing techniques and post-processing techniques (d'Alessandro et al., 2017). These approaches differ in terms of the point of the

AI lifecycle at which they intervene, the technical measures they employ and the circumstances in which they can be effectively applied. The table below presents an overview of techniques from each category and an illustrative application to the hypothetical recruitment system introduced above.

Table 3: Overview of bias mitigation measures

| Technique | Application to example | Preconditions and limitations | Suitable AI system |
|---|---|---|---|
| **Pre-processing techniques** aim at modifying datasets to better reflect diversity and reduce the risk that the model learns historical patterns of disadvantage. | | | |
| **Data provenance or datasheets** (Gebru et al., 2018; Holland et al., 2018). This is an **essential step before** other bias mitigation can be used. | Document how and why the recruitment dataset was created, who collected and curated it, what information it contains, and how and by whom it is intended to be used. It can identify historical intersectional gender imbalances in past hiring decisions before model training. It can include the listing of personal data (including gender) and non-personal data. | Relies on access to the dataset. | AI system agnostic, including GenAI |
| **Data statements** provide information on the source, linguistic and demographic characteristics of the data, the demographics of annotators, and the context in which the data were collected and used (Bender and Friedman, 2018). | A GenAI model is used to generate the job description. A data statement could clarify the source, linguistic and demographic characteristics of the training data, for example whether the model was primarily trained on corporate vacancy texts from particular industries, countries or workforce groups. It could further document annotator demographics, helping users understand whether the data were labelled or evaluated predominantly from particular cultural, linguistic or professional perspectives. | Access to the dataset.<br>Relies on information exchange and a shared understanding (of normative concepts) between providers and deployers. | Natural language processing, GenAI |
| **Causal analysis** to identify proxies. Identify and visualise relationships and dependencies between protected and non-protected variables (Glymour and Herington, 2019; Salimi et al., 2019). Identify and determine which dependencies are unfair. | Variables such as career interruptions, part-time work patterns, or membership of certain professional associations may be correlated with gender and indirectly influence the model's recommendations.<br>Effects of gender that could be identified as unfair and consequently not be allowed to influence the model's predictions:<br>**Gender → part-time work → hiring decision = potentially unfair pathway**<br>Management skills → hiring decision = likely legitimate and relevant pathway | Relies on access to the dataset.<br>Repairing the dataset – by removing dependencies based on pre-defined fairness metrics – requires contextual knowledge. | AI system agnostic, excluding GenAI |

| | | | |
|---|---|---|---|
| **Transformation of datasets** reduces the extent to which protected characteristics, or variables closely correlated with them, influence model predictions to meet specified fairness objectives (Calmon et al., 2017; Feldman et al., 2015; Zemel et al., 2013). | The model replaces the original (input) variables with transformed features that capture a candidate's overall suitability for the role, while limiting the influence of gender-related patterns reflected in part-time employment histories. The transformation seeks to **weaken the pathway 'gender → part-time work → hiring decision'**, while preserving the pathway 'management skills → hiring decision'. | Requires numerical data and preceding causal analysis.<br>Limitations include a trade-off between bias mitigation and performance; can make the model less interpretable. | AI system agnostic,<br>excluding GenAI |
| **Massaging or relabelling** is a type of transformation in which **labels** in the training data are modified to ensure equal distribution of positive outcomes (Kamiran and Calders, 2012).Can also be used for validation and testing datasets. | The label in the recruitment dataset is whether a candidate was hired **(Hired = Yes/No**). If historical hiring decisions systematically disadvantaged women, certain female candidates who were rejected despite having qualifications similar to successfully hired male candidates may have their **label changed from No to Yes.** Conversely, some positively labelled male candidates may be **relabelled from Yes to No** to restore balance. This reduces the extent to which the model learns biased historical hiring patterns. | Entails a trade-off between bias mitigation and performance. | AI system agnostic,<br>excluding GenAI |
| **Reweighing** (Calders et al., 2009; Jiang and Nachum, 2020; Krasanakis et al., 2018) means adding specific 'weight'[12] for data points to adjust for the bias in the training dataset.[13] | The recruitment dataset contains 1,000 candidates, of whom 50% are women. If 60% of all candidates are labelled as Hired = Yes, one would expect approximately 30% of the dataset to consist of women with a positive hiring label. Only 20% of the dataset consists of women labelled Hired = Yes. To adjust the bias, higher weights can be assigned to positively labelled female candidates and lower weights to positively labelled male candidates during model training. For example, a weight of 1.5 may be assigned to women with a positive hiring label and | Requires retraining the model on the modified dataset, which may be costly and resource-intensive for large-scale AI systems. | AI system agnostic,<br>excluding GenAI,<br>more suitable for smaller models |

[12] Weights are computed as the ratio between the expected and the observed probability of each group-outcome combination.

[13] The weights can be chosen based on three factors: (a) the special categories of personal data along with the probability in the population of this sensitive attribute; (b) the probability of a specific outcome [+/-]; and (c) observed probability of this outcome for a sensitive attribute.

| | | | |
|---|---|---|---|
| | a weight of 0.75 to men with a positive hiring label. | | |
| **Resampling** focuses on a balanced set of training samples by rebalancing the representation of different groups in the training data. Data from minority classes can be increased (oversampled), while data from the majority class can be decreased (undersampled) (Kamiran and Calders, 2012). | Women of colour constitute 15% of all applicants, but only 5% of the training dataset consists of women of colour labelled Hired = Yes, despite similarly qualified candidates from other groups being hired at higher rates. Oversampling can increase the representation of positively labelled women of colour from 5% to 15%, bringing it closer to their representation among applicants. Conversely, undersampling can reduce the representation of positively labelled candidates from overrepresented groups. | Undersampling requires datasets with sufficiently large amounts of data from different classes.<br>Effectiveness depends on the size of the entire dataset. | AI system agnostic,<br>excluding GenAI |
| **Generating artificial data**<br>(A part of) the dataset can be generated when the quantity of available data is limited (Sattigeri et al., 2019).[14] | Women of colour constitute 15% of all applicants, but only 5% of the training dataset consists of women of colour labelled Hired = Yes. Artificial or synthetic data generation can create additional positively labelled observations by interpolating existing candidates with similar characteristics. New artificial profiles may be generated that combine qualifications, management experience and employment histories commonly found among successfully hired women of colour. | Requires significant computing power.<br>Assumes an appropriate fairness criterion is available. | AI system agnostic,<br>especially suitable for unstructured data such as images |
| **In-processing techniques** seek to modify the model's behaviour. | | | |
| **Regularisation** incorporates an additional fairness objective into the model's performance metric. The model is trained to minimise both prediction error and | The model achieves high predictive accuracy but recommends interviews for 60% of qualified white male candidates and only 35% of equally qualified women of colour. This disparity contributes to the fairness violation term. During training, the optimisation process adjusts the model parameters to reduce this difference, for example by decreasing its reliance | Effectiveness depends on the chosen fairness metric (susceptible to gamification).<br>Trade-offs may arise between fairness, accuracy and robustness. | Supervised ML |

[14] The use of generative adversarial networks including specific bias considerations can contribute to generating biased datasets for training.

| | | | |
|---|---|---|---|
| fairness violations. The training objective: **Loss[15] = prediction error[16] + λ[17] × fairness violation.[18]** | on variables that indirectly reflect historical disadvantages. As a result, the model is encouraged to place greater reliance on job-relevant characteristics, such as management skills, professional experience and qualifications, rather than on gender-related employment patterns. | | |
| **Constrained optimisation** constrains the optimisation function by incorporating a fairness metric during the model training. In essence, this approach changes the algorithm of the AI model. | During training, the developer may require that equally qualified candidates from different groups, including women of colour, receive interview recommendations at similar rates. The model therefore seeks to maximise predictive performance while ensuring that historical underrepresentation among positively labelled candidates does not translate into unequal recommendation rates. If the disparity in interview recommendation rates between women of colour and other candidates exceeds a predefined threshold, the solution is rejected and the model parameters are adjusted until both performance and fairness requirements are met. | Effectiveness depends on the chosen fairness metric (susceptible to gamification).<br>Trade-offs may arise between fairness, accuracy and robustness. | Supervised ML |
| **Adversarial approach** means attacking the model to determine the protected attribute from the outputs (Celis and Keswani, 2019; Zhang et al., 2018). | An adversarial model attempts to predict whether a candidate is a woman of colour from the recruitment model's outputs. The adversary feedback can be used to penalise and update the model to prevent discriminatory outputs. | Computationally intensive, requires access to protected characteristics during training. | AI agnostic, including GenAI |
| **Post-processing techniques** seek to adjust model outputs after training to mitigate disparities in outcomes across groups. | | | |
| **Calibration** seeks to equalise the | The model may assign lower interview recommendation scores to candidates | Requires access to sensitive | Supervised ML |

[15] Loss refers to the overall objective function that the model seeks to minimise during training.

[16] Prediction error is the extent to which its predictions differ from the actual outcomes (e.g. incorrectly predicting which candidates are likely to be successful in the recruitment process).

[17] λ (lambda) is a tuning parameter that determines the relative importance assigned to fairness considerations compared with predictive accuracy during model training. A higher value of λ places greater weight on reducing disparities between groups, whereas a lower value prioritises predictive accuracy.

[18] Fairness violation refers to a measure of unfair disparities between groups (e.g. a substantial difference in positive hiring predictions between women of colour and other candidates with comparable qualifications).

| | | | |
|---|---|---|---|
| probability of positive outcomes across groups. It does not directly address the source of the bias. | with career breaks or women of colour. Calibration can adjust the model's output scores so that similarly qualified candidates across groups receive positive recommendations at more comparable rates. | attributes during deployment.<br>Involves trade-offs with predictive performance. | |
| **Thresholding** refers to adjusting decision thresholds after model training to reduce disparities in outcomes between groups. It assumes that biased outcomes often arise near the decision boundary, where small differences in scores can determine whether an individual receives an (un)favourable outcome. | The recruitment model recommends candidates for interview if their suitability score exceeds 0.70. A candidate from an under-represented group receiving a score of 0.69 may be moved above the decision threshold and treated as if they had received a score of 0.70 or higher, while a candidate from an overrepresented group receiving a score of 0.71 may be treated as falling below the threshold. The underlying model is not retrained; only the final classification decision is adjusted. | Some of the thresholding methods are not useful when there is historical bias. | Supervised ML |

Source: Selection and analysis of techniques by Shrishak (2024), adaptation with permission and reviewed by Shrishak.

The methods discussed above primarily target supervised learning systems that rely on labelled data. GenAI models are generally developed using self-supervised or unsupervised learning techniques, giving rise to the development of additional bias mitigation techniques (Shrishak, 2024). Although relevant, a discussion of these is beyond the scope of this study. While debiasing techniques specifically addressing non-binary gender bias remain relatively scarce compared with those targeting binary gender bias, a growing body of research has begun to develop that accounts for a broader spectrum of gender identities (Tronnier et al., 2024).

### 6.2.2. Fairness metrics

Researchers have developed methods for providing quantitative measures to assess unwanted bias and fairness in AI systems. These fairness metrics, sometimes also called bias tests, quantify differences in outcomes or performance across individuals or groups. In practice, fairness metrics are used both to identify potential fairness concerns and to evaluate whether a mitigation measure has been effective. Providers of high-risk systems are required to report evidence of performance gaps between protected groups and fairness metrics can help determine the severity of these gaps. Different metrics serve different purposes and answer different questions. The AI Act requires providers to comment on the appropriateness of the chosen performance metrics in the technical documentation of an AI system. Consider the hypothetical recruitment example introduced above. The AI system predicts whether a candidate should be recommended for an interview. Comparing these

predictions with the ground truth, that is, the candidate's actual suitability for the position, yields four possible outcomes: true positives (qualified candidates correctly recommended), false positives (unqualified candidates incorrectly recommended), true negatives (unqualified candidates correctly not recommended) and false negatives (qualified candidates incorrectly not recommended). Fairness metrics may assess whether these outcomes occur at similar rates across different groups. For example, they may examine whether women, including women of colour, experience higher rates of false negatives than men, which would mean that qualified female candidates are excluded from interview recommendations more often than qualified male candidates. Importantly, fairness metrics operationalise fairness in different ways. Depending on which metric is applied, the same system may appear fair or unfair with respect to different groups. No single fairness metric is appropriate in all circumstances. The choice of metric depends on factors such as the availability of agreed upon ground truth, the type of decision being made or facilitated by the system, and the nature of the harms associated with false positives and false negatives. The selection of a particular metric therefore determines which performance disparities become visible, and which remain hidden. Table 4 compares a number of commonly used group fairness metrics and the context in which their use may be justified.

Table 4: Sample comparison of group fairness metrics

| Fairness metrics | Description | When is the use of this metric appropriate? | Examples of AI applications | Harm addressed and mitigation |
|---|---|---|---|---|
| (Conditional) demographic parity | Different groups should receive positive outcomes at the same rate. | Situations where historic data is expected to be prejudicial, and there is no agreed-upon ground truth. | Hiring, offering loans, access to education, representation in the media. | Lack of selection; increase or decrease selection rate. |
| Equal opportunity | Different groups should experience false negatives at the same rate. | Situations where there is agreed-upon ground truth and the overwhelming harm comes from false negatives. | Cancer or other serious illness screening. | Failure to identify positive cases; increase the recall.[19] |
| Predictive parity | Different groups should have the same precision rate. | Situations where there is agreed-upon ground truth and the overwhelming harm comes from false positives. | Misidentification as a known person of interest to the police. | Failure to identify negative cases; increase the precision.[20] |
| False positive error rate balance | Different groups should receive false positives at the same rate. | Situations where there is agreed-upon ground truth and the overwhelming harm | Misidentification as a known person of interest to the police. | Failure to identify negative cases; increase the |

[19] Recall refers to the rate at which the model correctly predicts true positives.

[20] Precision refers to the rate at which the model correctly predicts true positives divided by the total number of positive predictions.

| | | | | |
|---|---|---|---|---|
| | | comes from false positives. | | true negative rate.[21] |

Source: Mittelstadt et al., 2024.

Risks of bias are often understood and measured using fairness metrics and then addressed using mitigation measures. However, the concept of 'risk of gender bias' is socially constructed, context-dependent and involves normative judgements regarding the likelihood, severity and acceptability of potential harms (Hacker et al., 2018). Under the AI Act's rules on risk management, measures must be implemented such that any 'relevant residual risk of the high-risk AI system' is judged to be 'acceptable'. This judgment sits with the AI system's provider. As such, academic literature has raised the question of whether providers of AI systems are best placed to identify and evaluate whether risks of bias are acceptable (Smuha and Yeung, 2025). As discussed in Chapter 5, the AI Act applies without prejudice to other Union legislation. While providers retain discretion in assessing and managing risks, AI systems remain subject to applicable anti-discrimination law, which constrains what may legally be regarded as an acceptable risk.

## 6.3. Organisational and governance measures to prevent and mitigate gender bias

Technological systems and social environment are mutually constitutive, and outcomes emerge through the dynamic interplay between social and technical elements. Therefore, one cannot be meaningfully studied in isolation from the other (Trist, 1981; Baxter and Sommerville, 2011). AI systems and models are embedded in complex organisational environments in which technical performance interacts with organisational and governance structures, cultural specificities and stakeholder expectations. As a result, many AI systems fail to become successfully implemented in real world settings (Kwarsick, 2026). Alongside technical bias mitigation measures, organisational and governance measures help identify the institutional norms, values and structural barriers that shape how AI systems operate in practice. AI governance can only become gender-responsive if it is grounded in the principles of inclusion, meaningful participation, accountability, substantive equality, transparency and oversight (UN Human Rights Council, 2026). Among the many governance measures that could be considered, this section focuses on two that directly use these principles: FRIAs and participation. Accountability and oversight mechanisms are discussed separately in Chapter 7. While most of the measures discussed in this chapter address gender bias as part of broader equality and fundamental rights governance frameworks, rather than through dedicated gender-responsive AI measures, they nevertheless offer valuable models for identifying and mitigating discriminatory risks throughout the AI lifecycle.

### 6.3.1. Fundamental rights and equality impact assessments, participation and prevention

The EU possesses a well-established body of gender mainstreaming practice, operationalised through gender impact assessments, reflected in national implementation across Member States, and further elaborated and codified through EIGE's step-by-step gender impact assessment methodology (EIGE, 2017). Yet, this existing framework has not been carried over into the EU digital acquis. Desk research

[21] The true negative rate refers to the rate at which the model correctly predicts true negatives.

and questionnaire responses at both EU and Member State level revealed no specific examples of gender impact assessment methodologies embedded in digital regulation. This gap points to a misalignment between the EU's long-standing gender mainstreaming practice and its digital regulatory framework. EIGE's methodology provides a tested ex ante model – including an initial assessment of whether a proposed measure, policy or system is likely to have gender-differentiated effects – analysing the current situation with sex-disaggregated data, weighing projected impacts and requiring gender expertise for the assessment. These elements could readily be integrated into fundamental rights assessments under the AI Act and the GDPR, as well as into systemic risk assessments under the DSA. The absence of such integration means that risks of algorithmic gender discrimination are currently addressed only incidentally through general equality provisions, rather than systematically through gender-structured assessment processes.

The analysis now turns to the AI Act, within which gender-related harms may be addressed through fundamental rights assessment. FRIAs are intended identify the specific risks to the rights of individuals – or groups of individuals – likely to be affected, and identify measures to be taken in the case of a materialisation of those risks. Article 27 of the AI Act requires certain deployers to conduct a FRIA prior to deployment of a high-risk AI system.[22] FRIAs are particularly valuable as a complement to the AI Act's risk management obligations for providers, as many discriminatory impacts only become apparent in the specific context in which an AI system is deployed (Birhane et al., 2022). Deployers are obliged to notify the results of the assessment to the MSA (Article 27(3)). The obligation, however, only applies to (i) public authorities, (ii) bodies governed by public law and private entities providing public services, as well as (iii) deployers of AI systems used for creditworthiness assessments, credit scoring, and risk assessment or pricing in life and health insurance (Article 27(1)). This limited scope has been criticised because the likelihood and severity of discriminatory impacts do not necessarily correspond to the AI Act's distinction between public and private deployers, nor to its categorisation of systems as high-risk. Discriminatory harms may also arise in contexts falling outside the scope of the FRIA obligation (Mantelero, 2024).

A second limitation that has been criticised is that the AI Act does not require the use of participatory methods in the context of the FRIAs[23] or other risk impact assessments. This is despite participatory technology development and assessment being considered to be important tools for identifying and addressing the social impacts of AI systems (UN Human Rights Council, 2011; DIHR, 2020; KULeuven et al., 2026). Participatory approaches are premised on the idea that affected groups, particularly vulnerable groups, should be included in the search for solutions to problems of social concern (Ahrweiler et al., 2025). Such involvement allows contextual and socio-cultural knowledge to complement technical approaches to bias detection and risk management (Shrishak, 2024). Participatory approaches may therefore improve the ability of impact assessments to identify harms, including gender-specific and intersectional, that might otherwise remain invisible to developers, deployers or assessors (ECNL and DIHR, 2025).

[22] Pursuant to Article 27 of the AI Act, the assessment shall consist of a description of the deployer's processes that includes: (a) the period of time within which and the frequency with which the AI system will be used; (b) the categories of natural persons and groups likely to be affected by use of the AI system in the specific context; (c) the specific risks of harm likely to impact those groups; (d) human oversight measures taken; and (e) and how those risks are addressed, including through internal governance and complaint mechanisms. At EU level, the AI Act embeds several of them in the FRIA.

[23] Recital 96 of the AI Act refers to the possiblity of including ('could involve') relevant stakeholders (including the representatives of groups of persons likely to be affected by the AI system), independent experts, and civil society organisations in conducting such impact assessments and designing measures to be taken in the case of materialisation of the risks.

The importance of participation is closely linked to broader concerns regarding representation in the design and governance of AI systems. As discussed in Chapter 3, women are under-represented among those who build and use AI. They make up around 30% of the AI workforce. Technological innovation has often been shaped by the perspectives of dominant social groups, leading to harms for women as well as their systematic exclusion from design and deployment processes (Criado-Perez, 2020). More diverse and inclusive teams cannot by themselves eliminate bias. They may reduce the risk that certain harms, assumptions, use cases or data gaps remain overlooked during the design, development and deployment of AI systems, however. Participation and diversity therefore operate as complementary safeguards: the former seeks to incorporate the experiences of affected groups, while the latter broadens the range of perspectives represented within the institutions developing and governing AI systems. At the same time, meaningful participation may itself be undermined by persistent gender inequalities in access to digital technologies, digital skills, technical education and leadership positions in the digital sector. The digital gender divide affects who can benefit from AI systems, and also who is able to shape, assess and govern them. Without concerted efforts to address the digital gender divide, the integration of AI into essential services and decision-making systems risks deepening existing inequalities across all areas of life (UN Human Rights Council, 2026).

Beyond their remedial function, FRIAs also serve an important preventive function. This function is reflected in a range of impact assessment methodologies, which require an assessment of whether a system is relevant, proportionate and appropriate to perform certain tasks before questions of risk mitigation arise (UNESCO, 2023; Government of Canada, 2025). In its Human Rights, Democracy, and the Rule of Law Impact Assessment (HUDERIA), the COE refers to these preliminary assessments as 'zero questions' (COE, 2025). The consideration of the suitability of the proposed AI system includes a reflection as to whether non-automated processes are considered.[24] As a result, an impact assessment may lead to the conclusion that an AI system should not be developed or deployed where fundamental rights risks cannot be adequately justified or mitigated. In the Netherlands, a FRIA conducted by the city of Rotterdam determined that a fraud detection system should not be used because of doubts about whether it could achieve the required level of transparency and legality (ENNHRI, 2023).

### 6.3.2. The use of AI to advance gender equality

While AI and related technologies hold genuine potential to promote gender equality and to begin addressing historical inequalities, a discussion of these opportunities falls outside the scope of this study. As the EU HLEG noted, 'even with good intentions, AI systems can cause unintentional harm'. As explained throughout this study, algorithmic discrimination is often the result not of deliberate exclusion but of design choices, data limitations or the reproduction of existing societal inequalities. Accordingly, one cannot presume that an AI system will produce beneficial outcomes simply because it is intended to advance gender equality.

The FRA provides a powerful illustration in its report Bias in Algorithms. The report examined NLP systems designed to detect offensive speech – a use case with clear relevance for gender equality – and found evidence of both gender bias and intersectional bias. For example, the Italian feminine term for Muslim, musulmana, led to a higher average in predicted probability of offensiveness (90%) than its masculine counterpart, musulmano (78%), and substantially higher score than the term cristiano (34%).

[24] In its *Ethical Impact Assessment. A Tool of the Recommendation on the Ethics of Artificial Intelligence*, UNESCO included Question 2.2.1 'Has careful consideration been given to non-algorithmic options which may be used to achieve the same goal? If so, why is the option involving an AI system favoured?'.

This suggests that the system associated particular negativity not only with religion but also with the intersection of religion and gender (FRA, 2022).

As the UN's Working Group on Discrimination against Women and Girls recently observed: 'While AI holds the potential to advance gender equality, such benefits will not be realized in the absence of deliberate, rights-based, feminist and gender-responsive regulation and governance' (UN, 2026).

## 6.4. Selection of best practices at European and national level

While most initiatives address gender bias as part of broader equality and fundamental rights governance frameworks, rather than through dedicated gender-sensitive and responsive measures, they nevertheless offer valuable models for identifying and mitigating discriminatory risks throughout the AI lifecycle.

### Practical tools for assessing (gender) discrimination in AI systems

- Within the framework of the EU Technical Support Instrument (2024-2026), implemented by the COE in cooperation with the Equality Bodies of Belgium, Finland and Portugal, a European assessment methodology on algorithmic discrimination was developed (EC, 2024b). The methodology provides a structured framework for investigating and assessing algorithmic discrimination cases.
- The EquiTech project (2024-2026), a project funded under the EU Citizens, Equality, Rights and Values (CERV) programme, the equality bodies of Estonia and Lithuania developed a set of practical tools for preventing and assessing algorithmic bias, including a bias impact assessment workbook, an algorithmic bias risk management methodology, and guidelines (The Gender Equality and Equal Treatment Commissioner, 2025).
- Equinet, together with Brent Mittelstadt, developed *How to Use the Artificial Intelligence Act to Investigate AI Bias and Discrimination: A Guide for Equality Bodies*. The guide provides practical guidance on how equality bodies can use the AI Act's transparency, documentation and oversight requirements when investigating discrimination complaints involving AI systems. It also advocates enforcement approaches that are more closely aligned with substantive rather than merely formal equality (Mittelstadt, 2024).
- The European Centre for Not-for-Profit Law and the Danish Institute Human Rights, working in cooperation with ECNL, developed a FRIA guide for deployers of high-risk AI systems under the AI Act. The guide is designed to support organisations in conducting FRIAs in line with the EU AI Act and relevant international standards (ECNL and DIHR, 2025).
- The Council of Europe HUDERIA methodology and model, adopted in 2025 and 2026, provide a non-binding framework for assessing and mitigating the risks and impacts of AI systems on human rights, democracy and the rule of law, supporting the implementation of the Council of Europe's AI governance framework (COE, 2025).
- The Dutch Algorithm Register enhances transparency regarding the deployment of algorithmic systems, and the Impact Assessment for Human Rights and Algorithms supports the systematic assessment of fundamental rights risks.

### Interdisciplinary, interinstitutional and multi-stakeholder cooperation

- In Spain, a formal cooperation framework to integrate gender equality throughout the entire cycle of AI supervision has been established between the Spanish Agency for the Supervision of AI, AESIA, and the Institute of Women and Equality Body (Ministerio de la Presidencia,

Justicia y Relaciones con las Cortes, 2026a). The agreement establishes joint work to ensure AI systems respect fundamental rights, mainly gender equality, by coordinating supervision, sharing technical expertise, and developing tools to detect and reduce discriminatory gender biases. It also includes collaboration on public awareness, training programs, and research initiatives such as 'idea labs' focused on ethical and inclusive AI. The two institutions exchange information, coordinate activities through a monitoring commission, and operate within their respective legal mandates for a four-year renewable term.

- France established PEReN (Pôle d'expertise de la régulation numérique), a multidisciplinary data science unit hosted by the Ministry of the Economy and made available to a range of public authorities and regulators. Through this arrangement, the French equality body, the Défenseur des Droits, can draw on specialised technical expertise in the investigation of algorithmic discrimination (Shrishak and Pénicaud, 2025).
- The Danish Institute for Human Rights (DIHR) positions stakeholder engagement as the core, cross-cutting component of its Human Rights Impact Assessment Guidance and Toolbox and provides concrete stakeholder engagement guidance (DIHR, 2020).
- To support the practical implementation of FRIA obligations, the AI Office is expected to launch an EU-wide pilot of a draft FRIA template before the end of 2026, with deployers invited to participate in the testing phase.

# 7. REDRESS, OVERSIGHT AND ACCOUNTABILITY MECHANISMS AND THEIR EFFECTIVENESS

KEY FINDINGS

- AI-related harms are often intangible and unnoticed, and detection frequently depends on journalists, researchers, equality bodies, and auditors. Proving AI-related harm requires technical-legal capacity that most individuals lack, making technical-legal aid increasingly essential to effective access to justice.
- Algorithmic gender-based discrimination is often intersectional and systemic, yet existing frameworks struggle to address it; monetary compensation cannot remedy collective and societal harm, which requires structural remedies and procedural rights with a societal dimension.
- Liability and access to redress are complicated by the diffusion of responsibility across the AI value chain, the transnational character of AI systems, and the fragmentation of the institutional landscape.
- The AI Act's complaint mechanism lacks the procedural safeguards and individual remedies of the GDPR. The DSA offers layered complaint and out-of-court mechanisms, but no comprehensive individual right to redress for platform failures.
- Collective redress provided under the equality directives, the Representative Actions Directive and the GDPR/DSA representation mechanisms are crucial, but cross-border obstacles persist.
- Oversight relies on expertise and cooperation from existing Fundamental Rights Authorities (AI Act) and trusted flaggers (DSA). However, these actors face resource constraints, which puts the sustainability of a governance model that relies on underfunded external actors into question.
- Self-assessed conformity with harmonised standards presumes compliance with the AI Act, giving providers strategic discretion that allows for potentially discriminatory outcomes, despite formal compliance. Similarly, DSA risk assessment largely delegates harm identification to the regulated companies themselves.
- With the AI Liability Directive withdrawn, no harmonised instrument addresses the evidentiary and causation challenges of algorithmic harm. The revised Product Liability Directive introduces disclosure obligations and rebuttable presumptions but excludes discrimination as such from actionable damage. Burden-of-proof rules in anti-discrimination law therefore remain critical. There is growing support for qualifying a failure to implement or document bias-mitigation measures as a prima facie case of discrimination, shifting the burden to the provider or deployer.

In most European systems, legal proceedings can be expensive, complex, and time-consuming (FRA, 2020). When addressing AI-related harms, pre-existing barriers to redress through legal proceedings are further exacerbated. The involvement of AI systems introduces additional layers of technical complexity and information and power asymmetries between affected individuals and providers or deployers of AI systems. AI-related harms often arise in contexts characterised by limited transparency, opacity, technical complexity and the involvement of multiple actors across the AI value chain, creating challenges for oversight, accountability, and redress (Dignum et al., 2025).

Understanding these challenges is essential for assessing whether existing legal and institutional mechanisms are capable of providing effective redress in practice. Building on the regulatory framework discussed in Chapter 5, Section 7.1 identifies four technology-specific barriers to effective redress to address AI- and ADM-related harms. In response to these challenges, the EU has strengthened its regulatory framework. Section 7.2 provides a non-exhaustive overview of the existing and emerging redress, oversight and liability mechanisms available in cases of algorithmic gender discrimination, and, to a lesser extent, TFGBV. This section evaluates the effectiveness of these mechanisms against the barriers identified in Section 7.1. The barriers to effective redress and legislative gaps identified in this chapter emerged as recurring concerns in the targeted consultation conducted for this study, where participants frequently highlighted obstacles to obtaining effective redress for AI- and ADM-related harms.

## 7.1. Barriers to effective redress for AI-related harm

Article 47 of the Charter of Fundamental Rights of the European Union enshrines the right to an effective remedy. The right, as interpreted by the CJEU and European Court of Human Rights (ECtHR), entails that avenues of redress must be practically and effectively accessible, not hindered by excessive informational, procedural, or financial obstacles, and available within a reasonable timeframe (FRA, 2020; Lenaerts, 2020). Remedies must be enforceable, meaning capable of providing adequate redress, whether through compensation, restitution, injunctive relief, or other measures appropriate to prevent, remedy or cease the violation in question (ECtHR, 2026; Peers et al., 2021). Four interrelated technology-specific barriers to effective redress to address AI- and ADM-related harms are discussed in this section: (1) barriers relating to the identification and reporting of harm; (2) barriers arising from the nature of the harm itself; (3) resource and expertise barriers; and (4) institutional and procedural barriers, including with a transnational dimension.

### 7.1.1. Barriers relating to the identification and reporting of harm

Effective redress presupposes that individuals are aware of the harm they have suffered. This is particularly challenging in the context of AI systems, as AI-related harms are often intangible and immaterial and may occur without the affected person's knowledge (Lázaro Cabrera and Maier, 2026). AI systems can, for example, exclude individuals from seeing advertisements for jobs, housing, or financial services, or offer more expensive products without individuals being aware of it (Wachter et al., 2021a). Other examples were discussed in Chapter 3, including being selected for investigation of fraud (section 3.2.1), paying higher insurance premiums or borrowing costs (section 3.2.2), or receiving a delayed diagnosis (section 3.2.3). Harm will often remain unnoticed given the opacity of the way in which AI systems are designed and operate. Even where individuals are aware of a potential harm, the harm may be perceived as insignificant, or in any case as too limited in proportion to costs, complexity and uncertainty associated with pursuing a complaint or legal action (Smuha, 2021).

The absence of relevant information may prevent individuals from identifying and contesting harms and thereby undermine the effective protection of fundamental rights. Transparency, including explainability, is widely regarded as a prerequisite for effective redress, as individuals can only challenge decisions, exercise their rights and seek remedies where sufficient information is available regarding the operation and effects of an AI system (Recital 72 and Article 86 AI Act; FRA, 2017; COE, 2024; COE 2026). Access to such information may, however, be limited by trade secret claims (Keller and Alpin, 2026), national security exemptions, which remain within the exclusive competence of the Member States pursuant to Article 4(2) of the TEU, and specific law enforcement exceptions. Notably, high-risk AI systems used in biometrics, law enforcement, and migration, asylum and border control management are registered only in the non-public part of the EU database (Article 49(4) AI Act).

Together, these limitations contribute to significant information asymmetries, consisting of imbalances in access to, understanding of and control over information (COE, 2024).

The importance of transparency for effective redress has also been confirmed in case law on algorithmic discrimination. In the *SyRI* case in the Netherlands, the District Court of The Hague attached significant weight to the absence of meaningful information regarding the functioning of the fraud-detection algorithm in finding the system incompatible with fundamental rights (Rechtbank Den Haag, 2020). Similarly, in the *Parcoursup* case, the French Constitutional Council recognised the right for student applicants to obtain information concerning the criteria, weighting and assessment methods used by the admission algorithm, introducing a transparency-by-default requirement for affected persons in the context of automated decision-making (Conseil Constitutionnel, 2020). See also Section 7.2.1.

Transparency alone does not fully resolve the problem of identifying harm, however. For effective redress, harm must also be detectable. In many cases, discriminatory effects only become visible through scrutiny of actors other than the affected individual. Investigative journalism, academic research, civil society organisations, equality bodies, auditors, and public authorities frequently play a crucial role in detecting algorithmic discrimination. Following investigative reporting on the gendered effects of personalised advertising systems, for example, the French Défenseur des Droits (2025) and the Dutch College voor de Rechten van de Mens (2025), both national equality bodies, concluded that Meta Ireland had engaged in indirect discrimination on grounds of gender (see also Chapter 2). Similarly, allegations following an analysis by investigative journalists of bias in a Dutch student-fraud detection system prompted an independent external audit, which ultimately confirmed the existence of indirect discrimination (Algorithm Audit, 2024).

Another barrier related to the identification of harm is the reporting barrier. Reporting barriers are particularly visible in the context of TFGBV. Victims frequently refrain from reporting due to stigma, fear of retaliation, concerns about re-victimisation and limited confidence that platforms or authorities will respond effectively (EIGE, 2026; FRA and EIGE, 2026). Those who do report directly to online platforms, for example to have a non-consensual intimate image or a hateful post removed, may find the notice-and-action mechanisms difficult to use, even though Article 16 DSA requires them to be easy to access and user-friendly (Rigotti, 2026a).

### 7.1.2. Barriers to redress arising from the intersectional, systemic, collective and societal nature of algorithmic harm

The mismatch between novel forms of harm arising through algorithmic AI and ADM and the existing legislative framework can leave AI-enabled harms outside the scope of traditional legal categories, and consequently leave victims without clear avenues for redress (Dignum et al., 2025; Pi and Proctor, 2025). In this section, this is illustrated through examples of such mismatches.

It is well established that EU gender equality and anti-discrimination laws, while offering important safeguards to address discrimination, also display inconsistencies, ambiguities, and shortcomings. These materialise especially when addressing algorithmic gender-based discrimination because of the nature of both gender-related harms and the nature of the technology. As discussed in Chapters 2, 3 and 5, gender-related harms are frequently intersectional in nature and existing legal frameworks continue to face difficulties in recognising, proving and remedying intersectional discrimination. Chapter 5 explains how novel manifestations of discrimination, such as proxy discrimination or discrimination based on grounds not recognised under existing legal frameworks, often arise in cases of gender bias but may not easily fit within established doctrinal categories (Wachter et al., 2021a; Gerards and Xenidis, 2021). The nature of algorithmic discrimination also challenges the traditional architecture of European and national anti-discrimination law. By relying on proxies and statistical

correlations rather than explicit classifications, algorithmic systems can obscure whether differential treatment is based directly on a protected characteristic or on a seemingly neutral criterion closely associated with that characteristic, thereby blurring the distinction between direct and indirect discrimination.

Similarly for victims of TFGBV, the harm combines dignity, privacy, and psychological dimensions that do not fit neatly within traditional categories of damage (EIGE, 2025; Rigotti, 2026a).

A further challenge is that much of the harm is systemic in nature. AI and ADM systems may affect individuals, groups and society simultaneously. As a result, they can give rise to individual, collective and societal harms (Smuha, 2021). Certain uses of AI systems risk exacerbating gender inequality, reinforcing misogynistic narratives and online communities, including those associated with the so-called manosphere, thus facilitating gender-based violence and the anti-gender backlash (EIGE, 2024; UN Women, 2023; Sandu and Cséby, 2026). As discussed in section 4.2.6, this can also manifest in the proliferation and normalisation of gender-based hate speech, which contributes to hostile online environments and the silencing of women and girls. By discouraging, silencing or constraining the participation of women in public life, they may also undermine democratic processes and pluralistic public debate (EIGE, 2024; UN, 2026). The resulting harms therefore extend beyond individual and group-based disadvantage, as they undermine gender equality as a societal interest and value.

Accordingly, remedies must be capable not only of repairing individual harm. While, monetary compensation can provide financial relief for individual harms, it is ill-suited to addressing collective and societal level harms. It is argued in academic literature that remedying such harms requires structural reform, such as mandatory redesign, (temporary) bans, compulsory external audits or product recalls (Wachter, 2024; Johnson et al., 2024). In addition, remedies should address procedural rights with a societal dimension, such as collective rights to access to information, access to justice, and participation in public decision-making on AI, irrespective of whether individual harm can be demonstrated (Smuha, 2021). Inspiration can be found in the Aarhus Convention on Access to Information, Public Participation in Decision-making and Access to Justice in Environmental Matters, where such rights are granted (Aarhus Convention, 1998).

The DSA requires VLOPs and VLOSEs to consider the severity and probability of systemic risks, with particular attention to the foreseeable negative impacts on fundamental rights including non-discrimination and harms related to gender-based violence. This legal qualification provides a framework for recognising and addressing the structural dimension of gendered harms, such as the over-moderation of consensual sexual content and amplifying misogynistic content. It moves beyond isolated incidents to consider the cumulative, societal impact of platform governance on sexual autonomy and gender equality (Rigotti, 2026a). However, the architecture of Article 34 of the DSA largely delegates risk identification to the regulated companies themselves, creating a high potential for corporate capture, where the interests of VLOPs and VLOSEs may unduly shape the assessment of the very harms the DSA seeks to prevent (Griffin, 2025).

### 7.1.3. Resource and expertise barriers

Individuals may face significant resource and expertise barriers when seeking redress for AI-related harms. Accessing and assessing relevant evidence, establishing causation, and meeting procedural requirements can be particularly difficult given the technical complexity of AI systems and the information and power asymmetries between affected individuals and AI providers or deployers (Gerards and Xenidis, 2021; Lázaro Cabrera and Maier, 2026). Given the difficulties associated with proving AI-related harms and establishing causation, data-driven investigations and algorithmic audits increasingly complement traditional forms of evidence gathering. Such investigations can help uncover

patterns of harm, generate evidence for regulatory oversight and inform strategic litigation concerning digital rights and algorithmic discrimination (AI Forensics, 2025; Chien et al., 2025; Digital Freedom Fund, 2023; Hooper et al., 2026). Article 47 of the Charter of Fundamental Rights of the European Union provides that 'legal aid shall be made available to those who lack sufficient resources in so far as such aid is necessary to ensure effective access to justice'. In the context of AI, effective access to justice may increasingly depend on forms of technical legal aid, capable of translating model documentation, impact assessments and audit reports into information that can be meaningfully used in legal proceedings (Lázaro Cabrera and Maier, 2026; Xenidis, 2025). It can be argued that effective access to justice does not only require legal assistance in the traditional sense but should extend to specialised technical expertise.

### 7.1.4. Institutional and procedural barriers, including transnational dimension

Effective redress for AI-related harms is hindered by the diffusion of responsibility across the AI value chain, by liability frameworks which are not well suited to address AI-specific harms, by transnational enforcement challenges, and by fragmented oversight (McGregor et al., 2019).

The large number of actors involved in the design, development, deployment and operation of AI systems can obscure responsibility and complicate the attribution of liability, a challenge discussed in the literature as the 'many hands problem' (Nissenbaum, 1996; Yeung, 2019; Coeckelbergh, 2020). While the AI Act allocates obligations across different actors in the AI value chain, including to providers, deployers, importers, and distributors, the translation of these responsibilities into effective avenues of redress remains challenging where harms result from the interaction of multiple actors and systems. These difficulties are compounded by the often-transnational nature of AI systems (UN Advisory Body on AI, 2024). AI-related harms frequently span multiple jurisdictions, as a single AI system might affect individuals across the EU.

These challenges are particularly visible in relation to TFGBV. A recent study on AI-generated NCII identified 88 publicly accessible websites hosting such content, including 38 dedicated exclusively to synthetic imagery, and found that many relied on services provided by a small number of major infrastructure providers. Rather than resting with a single actor, the creation, dissemination and accessibility of such content depended on a broader ecosystem involving AI developers, website operators, hosting providers, content delivery networks and search engines (Morgan et al., 2026). The findings provide a concrete illustration of the 'many hands problem' in digital environments, where harmful outcomes emerge from the interaction of multiple actors, each exercising only partial control over the relevant technologies and services. Although recent CJEU case law, including *Russmedia* (CJEU, 2025b) and *Coyote* (CJEU, 2026), suggests that intermediary status does not automatically preclude regulatory responsibility, the allocation and enforcement of legal obligations across this complex and transnational ecosystem remains challenging. At the same time, any expansion of the boundaries of intermediary liability requires careful calibration, as measures intended to strengthen accountability may also have implications for privacy, data protection, and other fundamental rights, particularly for individuals and groups already exposed to heightened risks of monitoring or surveillance. A notable example is the 2026 judgment of the Brussels Criminal Court, which held Twitter International Unlimited Company liable for failing to act expeditiously following repeated notifications of the non-consensual dissemination of intimate images (Correctional Court Brussels, 2026).

Pursuant to Article 4 TEU and Article 291 TFEU, enforcement and remedial mechanisms generally remain confined to national legal frameworks. Nevertheless, EU institutions and agencies have assumed an increasingly prominent role in the enforcement of EU law (Van Kreij, 2019; Scholten et al., 2024). Most relevant for this study are the enforcement powers exercised by the EC in the context of the DSA,

acting as MSA for VLOPs and VLOSEs pursuant to Article 65 DSA, and under the AI Act acting as MSA for GPAI models and very large platform-integrated systems following the recent omnibus amendments (Article 75a AI Act). While these developments may strengthen regulatory oversight at EU level, their relevance for securing effective redress for individuals who have suffered AI-related harms remains limited (see also 7.2.3).

Transnational corporate structures create additional barriers to accountability and jurisdictional disputes may further complicate access to redress (Leupold, 2022). In recent proceedings before European Equality Bodies concerning discriminatory advertising practices discussed above, Meta Ireland has argued that Irish law alone applies and has questioned the competence of equality bodies outside Ireland to assess the contested practices (Défenseur des droits, 2025; College voor de Rechten van de Mens, 2025). This difficulty is also present at the implementation level. In the ruling *Stitching Bits of Freedom v Meta*, the Amsterdam Civil Court of First Instance ordered Meta Ireland to provide users of Facebook and Instagram with the recommendation system of their choice, including a chronological, anti-profiled feed (Rechtbank Amsterdam, 2025). Although the ruling addressed a recommender system governed by the DSA and employed across the EU, its direct legal effects are limited to the Netherlands.

Finally, AI-related harms do not only span multiple jurisdictions, but they also cut across multiple regulatory regimes and supervisory mandates (Graux et al., 2025). Depending on the applicable legislation and the nature of the harm, affected individuals are confronted with a fragmented institutional landscape involving platform complaint mechanisms, data protection authorities, consumer protection authorities, (sectoral) market surveillance authorities, equality bodies, and civil and criminal justice systems. In certain Member States, this institutional complexity is further compounded by the coexistence of national and regional supervisory authorities, with overlapping or complementary mandates.[25]

## 7.2. Emerging rights and mechanisms addressing technology-specific barriers to redress

To address the challenges set out in Section 7.1, the EU has introduced several regulatory measures. Meanwhile, the framework emerged while a broader debate concerning innovation, competitiveness, and regulatory burden is ongoing (Draghi, 2024). Consequently, considerable emphasis is placed on self-assessment, conformity assessment procedures, risk management systems, and internal accountability structures within the framework (Wachter, 2024; Smuha and Yeung, 2025). The extent to which these mechanisms can effectively address technology-specific barriers to oversight, accountability, and redress remains to be seen, although academic and policy literature has been sceptical (Edwards and Veale, 2017; Lázaro Cabrera and Maier, 2026). A comprehensive overview of all remedies available to victims falls outside the scope of this study. This section therefore focuses on selected rights and mechanisms introduced at EU level to address technology-specific barriers to redress and examines their potential as well as their limitations. Given the size of the EU market and the regulatory influence often associated with the 'Brussels Effect', these rules may shape governance approaches beyond the EU (Bradford, 2020).

### 7.2.1. Information and transparency rights

The GDPR, the AI Act, and EU anti-discrimination laws contain rights and mechanisms for addressing information asymmetries and transparency barriers in cases of algorithmic gender discrimination.

[25] Similar institutional configurations can be observed across different regulatory domains, including equality and data protection law.

Several provisions in the GDPR and AI Act seek to strengthen the ability of affected individuals to identify, understand and challenge potentially discriminatory automated decision-making. Pursuant to Article 22 GDPR, individuals are protected against certain forms of automated decision-making. Articles 13, 14 and 15 of the GDPR grant individuals rights to information, including to meaningful information about the logic involved in ADM.[26] These rights have proven to be of particular importance for addressing algorithmic discrimination, including in situations where the AI Act and the rights and obligations set out therein are not applicable, for example because a system does not qualify as an AI system within the meaning of the AI Act or does not fall within the category of high-risk AI systems. In its *SCHUFA* judgment, the CJEU held that a credit score generated through profiling may itself classify as an automated decision within the meaning of Article 22 GDPR. The judgment offers a legal ground to challenge discrimination based on profiling and other decisions based overwhelmingly on automated risk scores (CJEU, 2023).[27] The Court of Justice further strengthened transparency rights in the *Dun and Bradstreet* judgment (CJEUa, 2025), holding that the right to obtain 'meaningful information about the logic involved' in accordance with Article 15(1)(h) GDPR must be understood as a right to an explanation of the procedure and principles actually applied in reaching a result. Such information 'must be provided by means of relevant information and in a concise, transparent, intelligible and easily accessible form'. At the same time, the Court acknowledged that transparency rights may need to be balanced against competing interests, including the protection of intellectual property and trade secrets. Importantly, the Court clarified that this balancing exercise must be carried out by a competent authority or court rather than by the company itself (Shrishak and Pénicaud, 2025). Because the scope of the GDPR is limited to identified or identifiable natural persons, it is less suited to address collective discrimination (see Section 7.2.4).

The AI Act complements these rights. Article 86 AI Act grants affected persons the right to obtain an explanation about the output of a high-risk AI system that produced legal or similarly significant effects to the health, safety, socio-economic, or any other fundamental rights. While this provision seeks to reduce information asymmetries in high-risk contexts, its scope remains limited to systems falling within the AI Act's high-risk regime, and providers have significant liberty to self-assess the applicability of the high-risk regime as outlined in Chapter 5. Pursuant to Article 49 AI Act, providers and deployers of AI systems are required to register information in the EU database (Article 71) on (i) High-risk AI systems listed in Annex III and (ii) AI systems which have been self-assessed by providers as not high-risk pursuant to Article 6(3). The information contained in the public part of the database should be 'accessible and publicly available in a user-friendly manner' (Article 71(4)). At the time of writing of this study, it is unclear when the EU database will be operational. Finally, Article 50 AI Act imposes transparency obligations on providers and deployers, requiring them to clearly inform natural persons when they deploy emotion recognition or biometric categorisation systems, systems that generate deepfakes, or AI-generated or manipulated text published to inform the public on matters of public interest.

Information asymmetries are also addressed through EU anti-discrimination law. The case law of the CJEU demonstrates that a lack of transparency (CJEU *Danfoss*, 1989) or refusal of information (CJEU *Meister*, 2012) can contribute to contextual evidence with a view to triggering a shift of the burden of proof (Weerts et al., 2023).

---

[26] The GDPR is currently subject to amendments under the Digital Omnibus. Amongst others, the EC proposes to limit data subject access requests in cases where they are made for reasons other than the protection of personal data. While it is not clear whether this amendment will survive negotiations in the EP and Council, it could have serious implications for data subjects wanting to access their data in cases of alleged discrimination or bias.

[27] Although the risk score provided by SCHUFA is transmitted to a third party, which is ultimately responsible for making the contractual decision, the Court argued that that third party 'draws strongly on that value' (Paragraph 48).

The DSA contains the most relevant transparency mechanisms in the context of TFGBV, including transparency obligations under Articles 14 (terms and conditions), 15 (transparency reporting for intermediary services), 17 (statement of reasons to affected persons in the context of content moderation), 24 (transparency reporting for platforms and search engines), 26 (recommender systems), 27 (advertising transparency) and 42 (transparency reporting for VLOPs and VLOSEs) as well as researcher access to platform data under Article 40. These mechanisms seek to facilitate the identification, documentation and investigation of harmful platform practices that may otherwise remain invisible to victims and regulators alike.

### 7.2.2. Complaint and individual redress mechanisms

The AI Act complements existing complaint and enforcement mechanisms that address algorithmic gender discrimination under equality and data protection law. Article 85 of the AI Act establishes the right to lodge a complaint for natural and legal persons with a MSA regarding infringements of the Act. Unlike its GDPR counterpart, however, this complaint mechanism is accompanied by only limited procedural safeguards. Whereas GDPR supervisory authorities are subject to specific obligations to handle and investigate complaints, the AI Act does not impose equivalent duties on MSAs. Nor does the AI Act provide equivalent individual remedies, such as the right to compensation (Article 82 GDPR), the right to rectification (Article 16 GDPR), or other data subject rights. As a result, the effectiveness of this complaint mechanism remains limited. Notable for this study is that the MSA is confined to assessing compliance with the obligations within the AI Act and as such does not assess potential discriminatory impacts. Importantly, this complaint mechanism is without prejudice to other existing remedies, such as filing a complaint with an equality body or DPA, or legal remedies that exist pursuant to national law, including directly going to court, assuming that the invoked provision of the AI Act has direct effect.

The DSA establishes several mechanisms to facilitate reporting, removal and contestation of harmful content, as well as to improve access to remedies for victims of TFGBV. The DSA establishes a layered system of platform complaint and redress mechanisms, which complement existing judicial and administrative avenues. For providers of intermediary services performing hosting functions, Article 16 of the DSA introduces a notice-and-action mechanism enabling users to notify providers of allegedly illegal content. Persons affected by a moderation decision must receive a statement of reasons to explain why their content was moderated (Article 17 DSA). The moderation may subsequently be challenged through internal complaint-handling mechanisms (Article 20 DSA) and certified out-of-court dispute settlement bodies (Article 21 DSA). These mechanisms seek to lower reporting barriers and provide accessible avenues for challenging platform decisions. The DSA does not establish a comprehensive individual right to redress for harms resulting from platform failures to remove illegal content or mitigate systemic risks (Leerssen et al., 2025; Rigotti, 2026a). The DSA preserves the right of recipients of services covered by the DSA to seek compensation for damage or loss suffered as a result of infringements of the DSA (Article 54 DSA). Remedies include both damages and injunctive relief, such as orders requiring the reinstatement of unlawfully removed content or user accounts (Ketteman and Tiedeke, 2020). Since obligations concerning systemic risk mitigation often leave providers considerable discretion as to the specific measures to be adopted to mitigate such risks, scholars have noted that this may limit the ability of private claimants to seek particular corrective actions through judicial proceedings (Leerssen et al., 2025). The extent to which individual provisions of the DSA can be enforced remains subject to ongoing legal debate.

Of particular relevance to this study is the relationship between the DSA and the VAW Directive. Article 23 of the VAW Directive provides for the removal of specific online content, referring to the principles and modalities of Article 9 DSA, while Article 42 promotes self-regulatory cooperation between

Member States and providers of intermediary services. Taken together, these provisions recognise that much of the abuse captured by the VAW Directive's criminal provisions, and potentially much that falls outside them, occurs on or through online platforms and search engines. More broadly, they show that TFGBV, and the VAW Directive's response to it, do not exist in a vacuum. Rather, they form part of a wider legal and regulatory landscape, including the other instruments examined in this study, which can complement the VAW Directive by offering victims protection through different branches of law and by targeting different stages of the lifecycle of TFGBV, from the design and deployment of technologies to the dissemination and moderation of harmful content (Rigotti et al., 2024).

### 7.2.3. Collective redress mechanisms

While not specifically designed to address collective and societal harms, collective enforcement and representative action mechanisms may play an important role where discriminatory practices cannot effectively be remedied through individual action alone. These mechanisms are particularly important in addressing information and power asymmetries between individuals and organisations, both in the context of discrimination as in the context of AI (De Gregorio and Demková, 2024).

Collective enforcement has long formed part of EU equality law. Articles 18(2) of the Employment Gender Equality Directive and 8(3) of the Gender Goods and Services Directive, as well as equivalent provisions in other EU equality directives, require Member States to ensure that associations, organisations, and other legal entities with a legitimate interest in enforcing equality law may engage, on behalf of or in support of a complainant and with their approval, in judicial and administrative proceedings. The CJEU further clarified in its judgment in the *Feryn* case that such entities may, in certain circumstances, bring proceedings in the absence of a specifically identified complainant, holding that discriminatory statements may constitute discrimination even in the absence of a specifically identifiable victim (*Feryn*, 2008). Moreover, in this ruling, the Court recognised that hypothetical harm may be sufficient to establish a presumption of discrimination. This is particularly relevant in the context of AI systems where affected individuals may not be identifiable, and discriminatory effects may only become visible at the group level.

The Representative Actions Directive (RAD), adopted by the EU in November 2020, introduced a framework for collective cross-sectoral consumer redress (EC, 2020). It is widely regarded as an important step for collective redress within the EU. Under this Directive, qualified entities can bring representative actions before national courts or administrative authorities on behalf of affected groups. These actions include seeking injunctive measures, for example to stop harmful practices from product or service providers, and to seek redress and compensation on behalf of 'affected persons'. The Directive does not provide the possibility of public interest litigation, despite its particular relevance in cases involving diffuse harms or unidentifiable victims. The RAD leaves EU Member States considerable discretion regarding its implementation. Member States may decide whether injunctive and redress measures can be pursued within the same procedure (Article 7(5) RAD). The RAD has been criticised for failing to address cross-border collective redress, insufficiently addressing current inadequacies of the legal framework designed for individual litigation in terms of jurisdiction (Brussels I-bis) and applicable law, or recognition and enforcement (Rome I and Rome II Regulations) (Leupold, 2022). Ireland remains the European headquarters for the majority of the largest US technology firms (Guinea et al., 2025) and consequently serves as a key jurisdiction for the enforcement of EU digital regulation. Ireland's domestic implementation of the RAD operates alongside strict historical domestic limitations on third-party litigation funding, limiting the potential for collective actions in Ireland due to the high costs associated with these proceedings (Guinea et al., 2025). The relevance of collective redress for AI and platform-related harms has been reinforced through the interaction between the RAD and the AI Act (Article 110 AI Act), as well as the DSA (RAD, Annex I (65)). This enables qualified entities to

bring representative actions in relation to infringements of these instruments. Such actions may be particularly valuable in situations where AI-related harms are collective and societal in nature and therefore difficult to address through individual complaints or litigation.

Whereas the RAD provides a dedicated framework for collective consumer redress, Articles 80 GDPR and 86 DSA establish representation mechanisms that are not confined to consumer protection. They allow non-profit organisations that meet specified criteria to exercise rights on behalf of data subjects under the GDPR and on behalf of recipients of intermediary services under the DSA, thereby facilitating collective enforcement and contributing to the effectiveness of the legislations.

### 7.2.4. Oversight and enforcement through and in cooperation with specialised actors

Both the AI Act and the DSA rely on cooperation between specialised actors to address harms that are difficult to detect, assess and remedy through traditional enforcement mechanisms alone.

To strengthen the oversight on discriminatory AI systems, the AI Act introduces cooperation mechanisms between MSAs and existing Fundamental Rights Authorities, so-called Article 77 bodies (see Section 5.1.1). Article 77 of the AI Act offers these Article 77 bodies new powers to effectively carry out their existing mandate, such as supervising anti-discrimination law, without prejudice to their existing competences, tasks and powers. These new powers include a right to access documentation (Article 77(1)) and a right to technical testing (Article 77(3)), as well as a right to cooperation from MSAs and AI operators in relation to all AI systems (Article 73(7), Article 79(1)(2) and Article 80(1)). These powers enable Article 77 bodies to contribute to the investigations of MSAs and potentially suggest relevant corrective actions, including withdrawal of the AI system from the market. These powers, when effectively exercised, can meaningfully contribute to assessing and remedying discrimination in AI systems. Following amendments introduced through the Digital Omnibus on AI, the information and testing rights can only be exercised through the MSA (new Article 77(1a)) (Regulation EU 2026/1744). As a result, the practical effectiveness of Article 77 depends on the resources of the MSA and Article 77 bodies, their willingness to cooperate and the concrete modalities governing cooperation between the authorities involved (Shrishak and Pénicaud, 2025; The Gender Equality and Equal Treatment Commissioner's Office of Estonia et al., 2024). To date, 22 Member States have identified equality bodies, including Gender Equality Commissions and National Human Rights Institutions, as Article 77 bodies.[28]

To strengthen the detection and moderation of harmful content, the DSA relies on cooperation between platforms, public authorities, civil society organisations and trusted flaggers. The DSA places the primary responsibility for moderation on platforms, retaining substantial discretion as to how harms are assessed and addressed. The regulatory mechanisms outsource part of the responsibility for content moderation from platforms to third party trusted flaggers (Article 22 DSA). Trusted flaggers identify and report illegal content, with platforms still making the ultimate decisions on removal. The effectiveness of these cooperation mechanisms relies on adequate resources for their functioning, including financial and human resources, that may not be available (Lázaro Cabrera and Maier, 2026). Emerging evidence suggests that both equality bodies, acting as Article 77 bodies in 22 Member States alongside other Article 77 bodies,[29] and CSO trusted flaggers (Sekwenz et al., 2026) face significant resource constraints and have not received dedicated funding corresponding with their expanding responsibilities. This raises questions regarding the sustainability of a governance model that

[28] See the European Commission's consolidated list of fundamental rights protection authorities designated under Article 77 AI Act, available at: https://digital-strategy.ec.europa.eu/en/policies/fundamental-rights-protection-authorities-ai-act.

[29] This observation is based on unpublished evidence gathered through member consultations conducted by the Equinet Secretariat and the ENNHRI Secretariat.

increasingly relies on so-far underfunded external actors to support the detection, assessment and remediation of technology-related harms (EIGE, 2025).

### 7.2.5. Accountability through self-assessment and risk management

As discussed in Chapter 6, self-assessment by providers of AI systems is the primary way that the AI Act expects implications for fundamental rights risks to be assessed. To certify compliance with the AI Act and its associated harmonised standards, conformity assessments must be undertaken (Article 43). Confirmation of compliance with the harmonised standards creates a presumption of compliance with the AI Act (Article 41), in effect giving the harmonised standards quasi-legal effect (Wachter; 2024). As a result, accountability under the AI Act is not only shaped by substantive legal obligations, but also by how providers define and operationalise requirements during the compliance process itself (Wachter, 2024; Smuha and Yeung, 2024). In the context of discriminatory AI systems, this discretion may be particularly consequential. For example, by defining broad intended purposes, low accuracy expectations for particular groups, or narrow ranges of potential misuses, providers may position a system to appear compliant with the AI Act while generating significant downstream harm when the system is deployed on the EU market. Lowering expected accuracy thresholds for historically under-represented groups, for instance, would render performance gaps between demographic groups consistent with the system's design specifications (Mittelstadt, 2024). Similarly, the intended purpose may restrict the scope of foreseeable unintended outcomes and risks, thereby functioning as a limitation on provider liability. In both cases, discriminatory outcomes may persist despite formal compliance with the AI Act.

The DSA introduces an *ex ante* governance framework aimed at identifying and mitigating systemic risks before they materialise. VLOPs and VLOSEs are subject to independent audit mechanisms, which assess compliance with their due diligence obligations including systemic risks (Article 37). However, the effectiveness of these audits is constrained by the indeterminate nature of the underlying obligations they assess. In particular, Article 34 DSA requires platforms to 'take into account' fundamental rights. However, it remains unclear whether this obligation is merely procedural, treating fundamental rights as one factor in internal risk assessments, or substantive, requiring outcomes to conform to fundamental rights standards. Consequently, the existence of an external audit mechanism does not resolve the underlying ambiguity of Article 34. In either case, the interpretation and application of equality rights are effectively delegated to private platforms, retaining substantial discretion in defining and weighing gender equality-related harms (Griffin, 2023; Karagianni and Calvi, 2026; EIGE, 2024). It is argued that the experience of technology-facilitated violence against women and girls – extensively set out in Chapter 4 – suggests that voluntary commitments and self-regulatory approaches are insufficient to prevent gendered harms at scale (EIGE, 2024).

### 7.2.6. Liability, causation and evidentiary burdens

Responsibility for AI systems that may be gender-biased is often diffused across developers, deployers, regulators, and data providers. This diffusion of responsibility complicates the attribution of liability and creates accountability gaps where no single actor can easily be held legally responsible for the harm caused. Limitations on holding a party liable for harm caused by an AI system are well established (Ebers, 2021; Hacker 2023). Following the withdrawal of the proposed AI Liability Directive, no harmonised EU instrument is in force or has been proposed that specifically addresses the evidentiary and causation challenges associated with AI-related harms. The revised Product Liability Directive (PLD) partially alleviates these difficulties by introducing disclosure obligations in Article 9 and rebuttable presumptions of defectiveness and causation in Article 10, including when the defendant fails to comply with disclosure requirements (Article 10(2)(c)). The PLD complements the AI

Act by providing a harmonised framework for non-contractual liability in cases involving defective AI-enabled products. However, its relevance for algorithmic discrimination remains limited. While the PLD in principle applies to both software- and AI-related products, the directive only covers material damages (Article 6(1)) and non-material damages (Article 6(2)) insofar they result from the material damages in Article 6(1) and in accordance with national law. Recital 24 specifies that damage consisting of pure economic loss, privacy infringements or discrimination do not by themselves trigger liability under the PLD. These are specifically harms that may arise from discriminatory AI systems. Faulty predictions can lead to dismissal, being denied loans, not being admitted to university, or being sent to prison – none of them qualifying as material harm in themselves, although material harm may arise from them. Recognising this, the EP recommended the incorporation of 'significant immaterial harm resulting in a verifiable economic loss' in the context of the – now withdrawn – AILD (EP, 2020). The limitation to material harm is particularly significant given that predictive and decision-making processes constitute some of the primary mechanisms through which discriminatory harms are produced and experienced (Wachter, 2024). Addressing non-material harms remains possible where this is foreseen in Member State law. The result is that AI-related civil liability in the EU reverts to the fault-based liability regimes of the individual Member States, with no harmonised mechanism to address the evidentiary difficulties specific to algorithmic harm, including discrimination (Wachter, 2024). Discrepancies between *ex ante* product safety obligations and *ex post* civil liability mechanisms can hinder effective redress for cumulative or distributed algorithmic harms (Arcila, 2024; Smuha and Yeung, 2025). Given all these barriers to redress arising from gender bias in AI systems via the EU digital acquis, the procedural safeguards in the anti-discrimination laws remain of critical importance.

Rules on the burden of proof are particularly important in ensuring that the principle of equal treatment can be effectively enforced. Article 19 of the Employment Gender Equality Directive and Article 9 of the Gender Goods and Services Directive, as well as equivalent provisions in the other EU equality directives, oblige Member States to provide for a shift in the burden of proof into domestic anti-discrimination regulations. This shift in the burden of proof postulates that the person who claims to have been discriminated against, or the representative entity acting on their or society's behalf, must first establish the facts from which discrimination may be presumed. If this is done successfully, the burden of proof shifts to the defendant, who must demonstrate that there has been no violation of the principle of equal treatment. This mechanism also applies in cases of algorithmic discrimination. Scholars have raised the question whether this principle suffices in the context of algorithmic discrimination, particularly given the limitations in terms of transparency and explainability. The idea that providers could be required to demonstrate that AI systems are free from discriminatory bias, particularly where adequate bias-mitigation measures have not been implemented, has received growing support in the literature (Wachter et al., 2021b).

As demonstrated in Chapters 2 and 3, historical bias is deeply embedded in sectors such as employment, finance, education, criminal justice, and healthcare, and the data collected to train and test AI reflect this bias. Rather than requiring the claimant to establish a presumption of discrimination directly, a failure to implement, document, or disclose adequate bias-mitigation measures could support a prima facie case of discrimination, thereby shifting the burden of proof to the provider or deployer to rebut that presumption (COE, 2026; Wachter, 2024; Xenidis, 2024). This argument could be reinforced by analogy with the CJEU's interpretation of preventive obligations under EU gender equality law in *Otero Ramos*, in the context of breastfeeding workers. The Court held that the 'failure to assess the risk posed by the work of a breastfeeding worker [...] must be regarded as less favourable treatment of a woman [...] and constitutes direct discrimination on grounds of sex' (CJEU, 2017). It further clarified that Article 4 of the Directive requires the employer to assess the risks and determine the protective and preventive measures that should be taken. Accordingly, it may be argued that

providers and deployers of AI systems are likewise subject to preventive obligations aimed at identifying and mitigating foreseeable risks to non-discrimination and gender equality. The significance of *Otero Ramos* lies in the Court's recognition that the failure to adequately assess risks and take preventive measures can itself qualify as direct discrimination. By analogy, the absence of appropriate bias assessment and mitigation measures may indicate that foreseeable risks of discrimination were not properly assessed and addressed throughout the AI lifecycle.

# 8. RECOMMENDATIONS

Within each thematic section, the recommendations are ordered based on their relative priority and feasibility.

**1. Strengthen the legal framework on gender equality**

- The EP is encouraged to renew its support for the Horizontal Equal Treatment Directive (COM(2008)426), and the Council is encouraged to reach unanimous agreement on this Directive, without further delay.
- The EU legislator and EU Member States should explicitly recognise intersectional discrimination in future legislation in line with Directive (EU) 2023/970, and existing legislation and policies should be interpreted and applied in a way that captures this.
- The EU legislator and EU Member States are expected to ensure appropriate legislation and policies are in place or are developed to prevent and combat all forms of discrimination throughout the lifecycle of AI systems.
- The EP and other relevant EU institutions are encouraged to support the development of interpretative guidance clarifying how existing EU non-discrimination law applies to algorithmic discrimination. This includes recognising proxy discrimination, facilitating access to evidence and ensuring that existing procedural safeguards are applied in a manner that effectively addresses the evidentiary challenges associated with algorithmic discrimination, including a presumption of algorithmic discrimination in case of failure to implement, document or disclose adequate bias-mitigation measures.
- The EP and other relevant EU institutions are encouraged to promote a gender-responsive interpretation and application of existing EU anti-discrimination legislation in the context of algorithmic decision-making.

**2. Strengthen accountability and access to justice for AI-related gendered harms**

The EU legislator is encouraged to consider the gaps and shortcomings identified in this study, by:

- Reintroducing the AI Liability Directive, covering material, immaterial, and societal harms and incorporating a strict liability regime, a robust reversal of the burden of proof, injunctive remedies and collective redress mechanisms.
- Designating additional high-risk systems exhibiting well-known and systemic risks via Articles 6(6), 7(1) and 97 of the AI Act, encompassing AI systems used in media, science and academia, most financial services and trading, and most types of insurance, as well as recommender systems, chatbots, and pricing algorithms. This recommendation is specifically directed to the EC.
- Recognising that effective access to justice in cases involving AI-related harms may require specialised technical expertise alongside traditional legal assistance, and encouraging Member States to take this need into account when applying existing legal aid and access-to-justice frameworks.
- Ensuring that efforts will be made towards the adoption of a future Directive on psychosocial risks, stress and mental health at work, which would address gendered AI risks in its protections against algorithmic management, so that algorithmic systems at work do not widen structural disadvantages for women, such as the gender pay gap and pension gap.

- Extending collective redress beyond Directive (EU) 2020/1828, by providing for public interest actions and procedural rights with a societal dimension, including access to information, access to justice and participation in public decision-making on AI, irrespective of demonstrated individual harm.

**3. Interpret and apply the digital acquis adopting a gender-responsive and intersectional approach**

- The EP is encouraged to connect EU-funded work on gender mainstreaming and gender impact assessments with the impact assessment methodologies in the digital acquis.
- Competent authorities are encouraged to verify that systemic risk assessments and independent audits of VLOPs and VLOSEs pursuant to the DSA explicitly address gender-based violence and discrimination against women arising from design choices.
- The EC and EU Member States are encouraged to frame AI governance around independent auditing, mandatory transparency, regulatory scrutiny and enforceable rights, rather than relying on self-regulation or voluntary, industry-led risk assessment. Gender equality should be treated not as an optional safeguard or compliance formality, but as a fundamental right and core value of the EU.
- National regulators and courts are encouraged to interpret and apply the EU digital acquis with due regard to gender sensitivity responsiveness. In other words, rather than treating its provisions as gender-neutral, they may actively identify and address the gender inequalities, including intersectional ones, that these provisions may produce or leave unaddressed.

**3. Ensure effective enforcement and oversight**

- EU Member States are expected to ensure that supervisory authorities responsible for enforcing the digital acquis have access to adequate financial, human and technical resources, as well as appropriate gender equality expertise, as a prerequisite for their functional independence and effective enforcement. EU Member States that have not yet complied with their obligations under Article 70 AI Act should designate their MSAs and single points of contact and notify the EC without delay.
- EU Member States are expected to designate Fundamental Rights Authorities with a gender equality mandate as being Article 77 bodies, pursuant to the AI Act. EU Member States and the EC are encouraged to ensure effective cooperation between MSAs and Fundamental Rights Authorities, including regarding access to the non-public part of the EU database (Article 49 AI Act) and adequate funding for these responsibilities, whilst monitoring the transposition of the directives on standards for equality bodies.
- The EC could consider mandating third-party conformity assessment for high-risk AI systems and GPAI models through delegated acts under Article 43(5)-(6) AI Act to close accountability gaps stemming from self-assessment.
- The EC may require that FRIAs under Article 27 AI Act are carried out in a participatory manner by persons with relevant expertise, and explicitly examine risks of gender-based discrimination, including from an intersectional perspective.
- The AI Office is encouraged to appoint experts on gender bias to its scientific panel.
- The European Centre for Algorithmic Transparency is encouraged to ensure that up-to-date methodologies for testing gender bias are accessible to public authorities, as established under Article 77 AI Act.

- The EC is encouraged to develop a code of conduct on gender bias and TFGBV pursuant to Directive (EU) 2024/1385 (Article 42), the DSA (Article 45) and the AI Act (Article 95), co-created with affected groups and their representatives.
- The EU legislator as well as the AI Office is encouraged to clarify, e.g. through guidelines, that transparency about bias-mitigation measures and accuracy testing does not affect trade secrets; source code need only be disclosed in exceptional cases, and only to supervisors.
- The EU and its Member States are encouraged to continue cooperating with international organisations to align responses to gender bias in AI and TFGBV, given that many of the relevant actors operate from outside the EU.

**4. Improve evidence collection, inclusive participation and funding of CSOs**

- The EP, the Council and the EU Member States are called upon to ensure consistent, structural and project-independent funding for CSOs within the next Multi-Annual Financial Framework. This should include funding for Trusted Flaggers under the DSA, centres against online abuse, and providers of legal and technical counsel as well as psychosocial support to victims.
- The EC and EU Member States are encouraged to support the collection of comparable, intersectional data on gender bias in AI and TFGBV, including through the regular surveys required by Article 44 of Directive (EU) 2024/1385.
- The EU legislator and Member States are encouraged to ensure that affected groups are meaningfully involved in the design and testing of AI systems by making such participation a mandatory element of FRIAs.

# REFERENCES


- Achterhold, E., Mühlböck, M., Steiber, N., & Kern, C. (2025). Fairness in Algorithmic Profiling: The AMAS Case. *Minds and Machines*, *35*(1), 9, https://doi.org/10.1007/s11023-024-09706-9.
- Adams, R., & Loideáin, N. N. (2019). Addressing indirect discrimination and gender stereotypes in AI virtual personal assistants: The role of international human rights law. *Cambridge International Law Journal*, *8*(2), 241–257, https://doi.org/10.4337/cilj.2019.02.04.
- Aguinis, H., Beltran, J. R., & Cope, A. (2024). How to use generative AI as a human resource management assistant. *Organizational Dynamics*, *53*(1), 101029, https://doi.org/10.1016/j.orgdyn.2024.101029.
- AI Forensics. (2025). *AI-generated image abuse: An update on Grok unleashed*, AI-Generated Image Abuse: An Update on Grok Unleashed | AI Forensics.
- Ahrweiler, P., Späth, E., Siqueiros García, J.M., Capellas, B.L., Wurster, D. (2025). Inclusive Technology Co-design for Participatory AI. In: Ahrweiler, P. (eds) Participatory Artificial Intelligence in Public Social Services. Artificial Intelligence, Simulation and Society. Springer, Cham, https://doi.org/10.1007/978-3-031-71678-2_2.
- AI HLEG. (2019a). *A definition of AI: Main capabilities and disciplines*, https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines.
- AI HLEG. (2019b). *Ethics Guidelines for Trustworthy AI*, https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html.
- Albiero, V., Zhang, K., King, M. C., & Bowyer, K. W. (2022). Gendered Differences in Face Recognition Accuracy Explained by Hairstyles, Makeup, and Facial Morphology. *IEEE Transactions on Information Forensics and Security*, *17*, 127–137, https://doi.org/10.1109/TIFS.2021.3135750.
- Aleiferis, M., Vlachoutsakos, P., & Kekesi, L. (2024, August 8). Χειροπέδες σε 51χρονο Σημαιοφόρο του Ναυτικού για εκδικητική πορνογραφία – Εκβίαζε 61χρονη με ερωτικό βίντεο. *Proto Thema*, https://www.protothema.gr/greece/article/1528267/heiropedes-se-51hrono-simaioforo-tou-nautikou-gia-ekdikitiki-pornografia-ekviaze-61hroni-me-erotiko-video.
- Allen, K. M., & Hermolle, M. (2026). 'This is going to be with me for the rest of my life': Perceived barriers to reporting technology-facilitated abuse. *Gender and Justice*, *2*(2), 143–168, https://doi.org/10.1332/30333660Y2026D000000034.
- Algorithm Watch (2022), "Automated decision-making systems and discrimination: a guidebook for anti-discrimination counselling". Available at: https://algorithmwatch. org/en/autocheck/.
- Aloisi, A., & De Stefano, V. (2021, January 7). Frankly, my rider, I don't give a damn. *Il Mulino*, https://www.rivistailmulino.it/a/frankly-my-rider-i-don-t-give-a-damn-1.
- Amnesty International. (2017). *Amnesty reveals alarming impact of online abuse against women*, https://www.amnesty.org/en/latest/press-release/2017/11/amnesty-reveals-alarming-impact-of-online-abuse-against-women/.
- Amnesty International. (2018). *Toxic Twitter—A toxic place for women*, https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1-1/.

- Amnesty International. (2024a). *Coded injustice: Surveillance and discrimination in Denmark's automated welfare state*, https://www.amnesty.org/en/documents/eur18/8709/2024/en/.
- Amnesty International. (2024b). *France: Discriminatory algorithm used by the social security agency must be stopped*, https://www.amnesty.org/en/latest/news/2024/10/france-discriminatory-algorithm-used-by-the-social-security-agency-must-be-stopped/.
- Amsterdam District Court 2 October 2025. Zaaknr. C/13/774725 / KG ZA 25-687 MK/JD), https://www.bitsoffreedom.nl/wp-content/uploads/2025/10/20251002-vonnis-kort-geding.pdf.
- Anderson, B. (2026). *Doxxed: How privacy abuse harms*. BRISTOL UNIVERSITY PRESS.
- Andraca, R. (2019, February 8). La Ligue du LOL a-t-elle vraiment existé et harcelé des féministes sur les réseaux sociaux ? Liberation, https://www.liberation.fr/checknews/2019/02/08/la-ligue-du-lol-a-t-elle-vraiment-existe-et-harcele-des-feministes-sur-les-reseaux-sociaux_1708185/.
- Andreeva, G., & Matuszyk, A. (2019). The Law of Equal Opportunities or Unintended Consequences?: The Effect of Unisex Risk Assessment in Consumer Credit. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *182*(4), 1287–1311, https://doi.org/10.1111/rssa.12494.
- Andrus, M., Spitzer, E., Brown, J., & Xiang, A. (2020). *'What We Can't Measure, We Can't Understand': Challenges to Demographic Data Procurement in the Pursuit of Fairness* (Version 2). arXiv, https://doi.org/10.48550/ARXIV.2011.02282.
- Are, C. (2022). The Shadowban Cycle: An autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies*, *22*(8), 2002–2019, https://doi.org/10.1080/14680777.2021.1928259.
- Association Belge des Consommateurs Test-Achats ASBL and Others v Conseil des ministres, Case C-236/09 (HvJ EU 1 March 2011), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62009CJ0236.
- Ayerza, J. M., Lee, J. R., O'Malley, R. L., & Lee, C. S. (2026). Cyberstalking and Online Dating: Examining Cyberstalking Victimization Among Dating App Users. *Asian Journal of Criminology*, *21*(1), 5, https://doi.org/10.1007/s11417-025-09480-2.
- Backe, E. L., Lilleston, P., & McCleary-Sills, J. (2018). Networked Individuals, Gendered Violence: A Literature Review of Cyberviolence. *Violence and Gender*, *5*(3), 135–146, https://doi.org/10.1089/vio.2017.0056.
- Barker, K., & Jurasz, O. (2020). Online violence against women as an obstacle to gender equality: A critical view from Europe. *European Equality Law Review*, *1*, 47–60.
- Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, *104*, 671–732, https://doi.org/10.15779/Z38BG31.
- Bates, S. (2017). Revenge Porn and Mental Health: A Qualitative Analysis of the Mental Health Effects of Revenge Porn on Female Survivors. *Feminist Criminology*, *12*(1), 22–42, https://doi.org/10.1177/1557085116654565.
- Bayerischen Rundfunk, & Report München. (2021). *Objective of biased: On the questionable use of AI for job applications*, https://interaktiv.br.de/ki-bewerbung/en/index.html.

- Baxter, G., & Sommerville, I. (2011). *Socio-technical systems: From design methods to systems engineering*. Interacting with Computers, 23(1), 4-17, https://doi.org/10.1016/j.intcom.2010.07.003.
- Begotti, T., Ghigo, M. A., & Acquadro Maran, D. (2022). Victims of known and unknown cyberstalkers: A questionnaire survey in an Italian sample. *International Journal of Environmental Research and Public Health*, *19*(8), 4883, https://doi.org/10.3390/ijerph19084883.
- Bender, E. M., & Friedman, B. (2018). *Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Transactions of the Association for Computational Linguistics*, https://doi.org/10.1162/tacl_a_00041.
- Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *The American Economic Review*, *94*(4), 991–1013.
- Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., & Caliskan, A. (2023). Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 1493–1504, https://doi.org/10.1145/3593013.3594095.
- Birhane, A, William, I, Prabhakaran, V, Díaz, M, Clare Elish, M, Gabriel, I, and Mohamed, S. (2022). *Power to the People? Opportunities and Challenges for Participatory AI. In Equity and Access in Algorithms, Mechanisms, and Optimization*, October 6–9, 2022, https://doi.org/10.1145/3551624.3555290.
- Blackwell, L., Chen, T., Schoenebeck, S., & Lampe, C. (2018). When Online Harassment Is Perceived as Justified. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1), https://doi.org/10.1609/icwsm.v12i1.15036.
- Bladini, M., & Svensson, E.-M. (2025). *Rethinking online anti-gender hate speech: Legislation and public policy making*. Routledge.
- Botero Arcila, B. (2024). *AI liability in Europe: How does it complement risk regulation and deal with the problem of human oversight?* Computer Law & Security Review, 54, 106012, https://doi.org/10.1016/j.clsr.2024.106012.
- Bouchaud, P. (2026). *Grok Unleashed*. AI Forensics, https://aiforensics.org/uploads/GrokUnleashed.pdf.
- Bouchaud, P., Çetin, R. B., Stanusch, N., Romano, S., & Faddoul, M. (2024). *Pay-to-play: Meta's community (double) standards on pornographic ads*. AI Forensics, https://aiforensics.org/uploads/report_meta_porn_ads.pdf.
- Bradford, A. (2020), *The Brussels Effect: how the European Union rules the world* (2020).
- Bujold, A., Roberge-Maltais, I., Parent-Rocheleau, X., Boasen, J., Sénécal, S., & Léger, P.-M. (2023). Responsible artificial intelligence in human resources management: A review of the empirical literature. *AI and Ethics*, https://doi.org/10.1007/s43681-023-00325-1.
- Bull, J. R., Rowland, S. P., Scherwitzl, E. B., Scherwitzl, R., Danielsson, K. G., & Harper, J. (2019). Real-world menstrual cycle characteristics of more than 600,000 menstrual cycles. *Npj Digital Medicine*, *2*(1), 83, https://doi.org/10.1038/s41746-019-0152-7.

- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, *81*, 77–91.
- Burgard, W. (2022). Artificial Intelligence: Key Technologies and Opportunities. In S. Voeneky, P. Kellmeyer, O. Mueller, & W. Burgard (Eds), *The Cambridge Handbook of Responsible Artificial Intelligence* (pp. 11–18). Cambridge University Press.
- Burgess, M. (2023, October 16). Deepfake is out of control. *Wired*, https://www.wired.com/story/deepfake-porn-is-out-of-control/.
- Butler, J. (2007). *Gender trouble: Feminism and the subversion of identity* (2nd ed). Routledge.
- Buyl, M., & De Bie, T. (2024). Inherent Limitations of AI Fairness. Communications of the ACM, 67(2), 48–55, https://doi.org/10.1145/3624700.
- Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). *Building Classifiers with Independency Constraints,* Building Classifiers with Independency Constraints | Proceedings of the 2009 IEEE International Conference on Data Mining Workshops.
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186, https://doi.org/10.1126/science.aal4230.
- Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized Pre Processing for Discrimination Prevention, https://doi.org/10.48550/arXiv.1704.03354.
- Celis, L. E., & Keswani, V. (2019). *Improved Adversarial Learning for Fair Classification,* https://doi.org/10.48550/arXiv.1901.10443.
- Centre for Intimacy Justice. (2026). *The digital gag: Suppression of sexual and reproductive health on Meta, TikTok, Amazon, and Google*, https://docsend.com/view/emzyirq6hfatmx2a.
- Centre Hubertine Auclert. (2018). *Rapport sur les cyberviolences conjugales*, https://www.centre-hubertine-auclert.fr/egalitheque/publication/rapport-cyberviolences-conjugales.
- Chayn. (2026). *Explicit harms of non-explicit images: Defining image-based abuse in Pakistan and diaspora*, https://reports.chayn.co/iba_pdfs/chayn-report-explicit-harms-of-non-explicit-images.pdf.
- Chien, C. V., Zink, A., & Chen, I. Y. (2025). *Bridging research gaps between academic research and legal investigations of algorithmic discrimination* (arXiv:2508.14954). arXiv, https://arxiv.org/abs/2508.14954.
- Chopin, I., & Germaine, C. (2026). A comparative analysis of non-discrimination law in Europe 2025: The 27 EU Member States compared. Publications Office of the European Union, https://doi.org/10.2838/6732664.
- Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Gigante, A., Valencia, A., Rementeria, M. J., Chadha, A. S., & Mavridis, N. (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *Npj Digital Medicine*, *3*(1), 81, https://doi.org/10.1038/s41746-020-0288-5.
- Citron, D. K. (2009). Law's Expressive Value in Combating Cyber Gender Harassment. *Michigan Law Review*, *108*(3), 373–416.

- Citron, D. K. (2023). The Continued (In)visibility of Cyber Gender Abuse. *The Yale Law Journal Forum*, 333–371.
- Coeckelbergh, M. (2020). *Artificial intelligence, responsibility attribution, and a relational justification of explainability*. *Science and Engineering Ethics, 26*(4), 2051–2068, https://doi.org/10.1007/s11948-019-00146-8.
- Cole, S. (2018, January 24). We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now. *Vice,* https://www.vice.com/en/article/reddit-fake-porn-app-daisy-ridley/.
- College voor de Rechten van de Mens. (2025). *Meta Platforms Ireland Ltd. discriminates on the ground of gender when displaying job advertisements to users of Facebook in the Netherlands* (Oordeel 2025-17), https://oordelen.mensenrechten.nl/oordeel/2025-17/4a575c22-d4b0-499f-8811-6b5e6720344d.
- Committee on the Elimination of Discrimination against Women. (1992). *General recommendation No. 19: Violence against women* (A/47/38). United Nations, https://undocs.org/A/47/38.
- Committee on the Elimination of Discrimination against Women. (2017). *General recommendation No. 35 on gender-based violence against women, updating general recommendation No. 19* (CEDAW/C/GC/35). United Nations, https://undocs.org/CEDAW/C/GC/35.
- Conseil constitutionnel, Décision n°2020-834QPC du 3 avril 2020, Décision n° 2020-834 QPC du 3 avril 2020, Union nationale des étudiants de France [Communicabilité et publicité des algorithmes mis en œuvre par les établissements d'enseignement supérieur pour l'examen des demandes d'inscription en premier cycle].
- Corbin, B. A. (2020). Digital micro-aggressions and discrimination: Femtech and the 'othering' of women. *Nova Law Review*, *44*, 337–364.
- Correctional Court Brussels. (2026, April 7). Twitter International Unlimited Company (Case No. 23N003625). Institute for the Equality of Women and Men, 2026-04-7 Correctionele rechtbank Brussel.pdf.
- Council Directive 2000/43/EC of 29 June 2000 Implementing the Principle of Equal Treatment between Persons Irrespective of Racial or Ethnic Origin, CONSIL, 180 OJ L (2000), http://data.europa.eu/eli/dir/2000/43/oj.
- Council of the European Union Decision (EU) 2026/1080 of 21 April 2026 on the conclusion, on behalf of the European Union, of the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law.
- Council of Europe. (2024). *Explanatory report to the Council of Europe Framework Convention on*
- *Artificial Intelligence and Human Rights, Democracy and the Rule of Law* (CETS No. 225). Council of Europe. https://rm.coe.int/1680afae67.
- Council of Europe, European Court of Human Rights. (2025). *Guide on Article 13 of the European*
- *Convention on Human Rights: Right to an effective remedy.* https://ks.echr.coe.int/documents/d/echr-ks/guide_art_13_eng; Peers, S., Hervey, T., Kenner, J., & Ward, A. (Eds.). (2021). The EU Charter of Fundamental Rights: A Commentary (2nd ed.). Hart Publishing.
- Council of Europe. (2025). *HUDERIA Methodology and model,* April 2026.

- Council of Europe. (2026). *European assessment methodology on algorithmic discrimination.* https://rm.coe.int/european-assessment-methodology-on-algorithmic-discrimination/48802dac82.
- Court of Justice of the European Union (2008) S. Coleman v Attridge Law and Steve Law, EU:C:2008:415.
- Court of Justice of the European Union (2023), C-634/21 SCHUFA Holding (Scoring) EU:C:2023:957.
- Court of Justice of the European Union (2025a), C-203/22 Dun & Bradstreet Austria EU:C:2025:117.
- Court of Justice of the European Union (2025b), C-492/23 Russmedia Digital SRL ECLI:EU:C:2025:935.
- Court of Justice of the European Union (2026), C-188/24 WebGroup Czech Republic and NKL Associates and C-190/24 Coyote System, ECLI:EU:C:2026:492.
- Court of Justice of the European Union (2018) Centrum voor gelijkheid van kansen en voor racismebestrijding v. Firma Feryn NV, Case C-54/07, ECLI:EU:C:2008:397.
- Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *The University of Chicago Legal Forum*, *1*(8), 139–167.
- Criado-Perez, C. (2020). *Invisible women: Exposing data bias in a world designed for men*. Vintage.
- Cui, C., Wu, Y., & Zhang, Y. (2026). *From Celebrities to Anyone: Characterizing AI Nudification Content, Technology, and Community Dynamics on 4chan* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2606.27234.
- Daffalla, A., Chao, S., & Zeng, E. (2026). *Dual-Use AI Face Swap Apps Are Mostly Unsafe: A Systematic Safety Audit* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2605.24735.
- Danaher, J. (2018). The law and ethics of virtual sexual assault. In W. Barfield & M. Blitz, *Research Handbook on the Law of Virtual and Augmented Reality* (pp. 363–388). Edward Elgar Publishing. https://doi.org/10.4337/9781786438591.00021.
- Danks, D., & London, A. J. (2017). Algorithmic Bias in Autonomous Systems. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 4691–4697. https://doi.org/10.24963/ijcai.2017/654.
- Dastin, J. (2018, October 11). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.
- De Gregorio, G. and Demková, S. (2024). *The constitutional right to an effective remedy in the digital age: a perspective from Europe.* In European Yearbook of Constitutional Law 2023: Constitutional Law in the Digital Era (pp. 223-254). The Hague: TMC Asser Press.
- Défenseur des droits. (2025). *Decision No. 2025-182*. Défenseur des droits.
- Devuyst, Y. (2025). The EU's Horizontal Equal Treatment Directive: A Way Forward (SIEPS).
- Digibeetle. (2026) *Code Red – Supervisory Authorities: Understanding Their Key Challenges*, September 2026 (v.4), https://digibeetle.eu/full-sa-report/.

- Digital Freedom Fund. (2023). *Evidence gathering for platform accountability* [Report]. Digital Freedom Fund. https://personaldata.io/wp-content/uploads/2024/11/DFF-Evidence-Gathering-Report-v3_compressed.pdf.
- Dignum, V., Carli, R., & Tang, Y. (2025). *Background paper: Human rights & artificial intelligence. The 23rd Informal ASEM Seminar on Human Rights.* Asia-Europe Foundation (ASEF). https://asef.org/wp-content/uploads/2025/10/ASEMHRS23-Backgound-Paper.pdf.
- DIHR, Danish Institute for Human Rights. (2020). Cross-cutting: Stakeholder engagement. Human rights impact assessment guidance and toolbox (Report A4). https://www.humanrights.dk/files/media/document/HRIA%20Toolbox_Stakeholder%20Engagement_ENG_2020.pdf.
- Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the Implementation of the Principle of Equal Opportunities and Equal Treatment of Men and Women in Matters of Employment and Occupation (Recast), EP, CONSIL, 204 OJ L (2006). http://data.europa.eu/eli/dir/2006/54/oj.
- Doudesis, D., Lee, K. K., Boeddinghaus, J., Bularga, A., Ferry, A. V., Tuck, C., Lowry, M. T. H., Lopez-Ayala, P., Nestelberger, T., Koechlin, L., Bernabeu, M. O., Neubeck, L., Anand, A., Schulz, K., Apple, F. S., Parsonage, W., Greenslade, J. H., Cullen, L., Pickering, J. W., … Smith, S. W. (2023). Machine learning for diagnosis of myocardial infarction using cardiac troponin concentrations. *Nature Medicine*, *29*(5), 1201–1210. https://doi.org/10.1038/s41591-023-02325-4.
- Draghi, M. (2024). *The future of European competitiveness: A competitiveness strategy for Europe* (Part A). European Commission. https://commission.europa.eu/topics/competitiveness/draghi-report_en.
- Dragiewicz, M., Burgess, J., Matamoros-Fernández, A., Salter, M., Suzor, N. P., Woodlock, D., & Harris, B. (2018). Technology facilitated coercive control: Domestic violence and the competing roles of digital media platforms. *Feminist Media Studies*, *18*(4), 609–625. https://doi.org/10.1080/14680777.2018.1447341.
- Duane, M., Contreras, A., Jensen, E. T., & White, A. (2016). The Performance of Fertility Awareness-based Method Apps Marketed to Avoid Pregnancy. *The Journal of the American Board of Family Medicine*, *29*(4), 508–511. https://doi.org/10.3122/jabfm.2016.04.160022.
- Duffy, B. E., & Schwartz, B. (2018). Digital 'women's work?': Job recruitment ads and the feminization of social media employment. *New Media & Society*, *20*(8), 2972–2989. https://doi.org/10.1177/1461444817738237.
- Dunn, S. (2020). *Technology-Facilitated Gender-Based Violence: An Overview*. Centre for International Governance Innovation. https://www.cigionline.org/publications/technology-facilitated-gender-based-violence-overview/.
- Dunn, S. (2021). Is it actually violence? Framing technology-facilitated abuse as violence. In J. S. Bailey, A. Flynn, & N. Henry (Eds), *The Emerald international handbook of technology-facilitated violence and abuse* (First edition, pp. 25–45). Emerald Publishing.
- Dunn, S. (2024). Legal definitions of intimate images in the age of sexual deepfakes and generative AI. *McGill Law Journal*, *69*(4), 396–416.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. S. (2012). *Fairness through awareness.* https://doi.org/10.48550/arXiv.1104.3913.

- d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). *Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification*. https://doi.org/10.48550/arXiv.1907.09013.
- d'Ursel, E. (2023). Accession to the Convention. In S. De Vido & M. Frulli (Eds), *Preventing and Combating Violence Against Women and Domestic Violence* (pp. 804–820). Edward Elgar Publishing. https://doi.org/10.4337/9781839107757.ch83.
- Algorithm Audit (2024), Vervolgonderzoek bevestigt indirecte discriminatie controles uitwonendenbeurs, via https://duo.nl/particulier/home/actueel/vervolgonderzoek-bevestigt-indirecte-discriminatie-controles-uitwonendenbeurs.jsp.
- Earle, S., Marston, H. R., Hadley, R., & Banks, D. (2021). Use of menstruation and fertility app trackers: A scoping review of the evidence. *BMJ Sexual & Reproductive Health*, *47*(2), 90–101. https://doi.org/10.1136/bmjsrh-2019-200488.
- Ebers, M. (2021). *Liability for Artificial Intelligence and EU Consumer Law*, 12 J. INTELL. PROP. INFO. TECH. & ELEC. COM. L. 204, 216-19 (2021).
- Economist Intelligence Unit. (2021). *Measuring the Prevalence of Online Violence against Women*. https://onlineviolencewomen.eiu.com.
- Edelenbos, L., Kort Illeborg, C., Møller Mundt, C., & Bødtker, M. (2026). *En livslang regning: En analyse om konsekvenserne af økonomisk partnervold mod kvinder*. Danner. https://danner.dk/wp-content/uploads/2026/02/En-livslang-regning_En-analyse-om-konsekvenserne-af-oekonomisk-partnervold-mod-kvinder_Danner_2026.pdf?ref=fullerproject.org.
- Edwards, L. and Veale, M. (2017*). Slave to the algorithm? why a right to an explanation is probably not the remedy you are looking for.* Duke L. & Tech. Rev., 16:18, 2017.
- European Data Protection Supervisor (EDPS) & European Data Protection Board (EDPB) (2021). Joint Opinion 5/2021.
- EIGE. (2021). *The costs of gender-based violence in the European Union*. https://eige.europa.eu/news/gender-basedviolence- costs-eu-eu366-billion-year.
- EIGE. (2022). *Cyber Violence against Women and Girls Key Terms and Concepts*. https://eige.europa.eu/sites/default/files/cyber_violence_against_women_and_girls_key_terms_and_concepts.pdf.
- EIGE. (2024). T*ackling cyber violence against women and girls: The role of digital platforms Publications Office*. https://doi.org/10.2839/1955989.
- EIGE. (2025). *Impact driver: Marking milestones and opportunities for gender equality in the EU.* Publications Office. https://data.europa.eu/doi/10.2839/8280076.
- EIGE. (2026). *From lived reality to policy action*. Publications Office of the European Union. https://doi.org/10.2839/5514733.
- EIGE. (2017). Guide to gender impact assessment. https://eige.europa.eu/gender-mainstreaming/toolkits/gender-impact-assessment/guide-gender-impact-assessment.
- Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2017). *Runaway Feedback Loops in Predictive Policing* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1706.09847.

- Equinet (2024), *How to use the Artificial Intelligence Act to investigate AI bias and discrimination: A guide for Equality Bodies*.
- Equinet, & European Network of National Human Rights Institutions (ENNHRI). (2026, March). *Equinet*.
- *ENNHRI statement on the Digital Omnibus Regulation proposals on AI and on data*. https://equineteurope.org/wp-content/uploads/2026/03/Equinet-and-ENNHRI-joint-statement.pdf.Eticas Foundation. (2022). *The external audit of the VioGén system: Spain's algorithm designed to protect victims of gender violence*. https://dev.eticasfoundation.org/gender/the-external-audit-of-the-viogen-system/.
- Equitech (2026). *The Gender Equality and Equal Treatment Commissioner*. *EquiTech project*. https://www.volinik.ee/en/node/1823.html.
- European Banking Authority. (2025). *Rising application of AI in EU banking and payments sector*. https://www.eba.europa.eu/sites/default/files/2025-09/146b3558-d026-47bf-a872-f05e93ed30d2/Rising%20application%20of%20AI%20in%20EU%20banking%20and%20payments%20sector.pdf.
- European Center for Not-for-Profit Law, & Danish Institute for Human Rights. (2025). *A guide to fundamental rights impact assessments (FRIA) under the EU Artificial Intelligence Act*. https://www.humanrights.dk/files/media/document/ECNL%20DIHR%20Guide%20to%20FRIA%20Dec%202025.pdf.
- European Commission. (2012). *Guidelines on the application of Council Directive 2004/113/EC to insurance, in the light of the judgment of the Court of Justice of the European Union in Case C-236/09 (Test-Achats)*. https://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2012:011:0001:0011:en:PDF.
- European Commission. (2018). *Artificial Intelligence for Europe*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN.
- European Commission. (2020a). *Legal gender recognition in the EU: The journeys of trans people towards full equality*. Publications Office. https://doi.org/10.2838/085419.
- European Commission. (2020b). *White paper on artificial intelligence—A European approach to excellence and trust*. https://commission.europa.eu/document/download/d2ec4039-c5be-423a-81ef-b9e44e79825b_en?filename=commission-white-paper-artificial-intelligence-feb2020_en.pdf.
- European Commission. (2022). Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive), COM(2022) 496 final. EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0496.
- European Commission. (2023). *Commission implementing decision of 22.5.2023 on on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in support of Union policy on artificial intelligence*. https://ec.europa.eu/transparency/documents-register/detail?ref=C(2023)3215&lang=en.

- European Commission. (2024a) Directorate General for Justice and Consumers. *Gender stereotypes: Violence against women: Eurobarometer report.* Publications Office. https://doi.org/10.2838/982236.
- European Commission. (2024b). *Upholding equality and non-discrimination by Equality bodies regarding the use of artificial intelligence in public administrations. Reforms and Investments.* https://reforms-investments.ec.europa.eu/technical-support-instrument-0/labour-market-and-social-protection/upholding-equality-and-non-discrimination-equality-bodies-regarding-use-artificial-intelligence_en.
- European Commission. (2025a). *Commission Guidelines on the definition of an artificial intelligence system established by Regulation (EU) 2024/1689 (AI Act)*.
- European Commission (2025b), *Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act)*, *(42) and (43)* available at https://digital-strategy.ec.europa, accessed 14 August 2026.
- European Commission. (2025c). European Commission work programme 2025, *Moving forward together: A bolder, simpler, faster Union* COM(2025) 45, 11.2.2025, Annex IV: Proposals to be withdrawn, No 26. https://commission.europa.eu/.
- European Commission. (2025d). *A comparative analysis of non discrimination law in Europe 2025*, https://www.equalitylaw.eu/downloads/6477-a-comparative-analysis-of-non-discrimination-law-in-europe-2025.
- European Commission. (n.d.). *An agile Digital Rulebook for the EU*. Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/policies/digital-rulebook.
- European Commission (2026), *Draft Commission guidelines on the classification of high-risk AI systems under Article 6 of Regulation (EU) 2024/1689 (AI Act) for stakeholder consultation*, *(169).*
- European Court of Justice (109/88). (1989). Handels- og Kontorfunktionærernes Forbund I Danmark v Dansk Arbejdsgiverforening, acting on behalf of Danfoss. EU:C:1989:383.
- European Court of Justice (C-415/10), Meister, G. v. Speech Design Carrier Systems GmbH, Case C-415/10, ECLI:EU:C:2012:217.
- European Data Protection Board (2021). *Overview on resources made available by Member States to the Data Protection Authorities and on enforcement actions by the Data Protection Authorities*; 82% of the authorities reported that existing human and financial resources were insufficient to carry out their tasks effectively.
- European Data Protection Board [EDPB], & European Data Protection Supervisor [EDPS]. (2026). *EDPB-EDPS Joint Opinion 1/2026 on the Proposal for a Regulation as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI)*. European Data Protection Board. https://www.edpb.europa.eu/documents/legislative-opinion/edpb-edps-joint-opinion-12026-on-the-proposal-for-a-regulation-as.
- European Institute for Gender Equality. (2021). *Artificial intelligence, platform work and gender equality.* Publications Office. https://doi.org/10.2839/805305.
- European Institute for Gender Equality (Ed.). (2025). *Combating cyber violence against women and girls: Developing an EU measurement framework: methodological report*. Publications Office. https://doi.org/10.2839/5652503.

- European Insurance and Occupational Pension Authority. (2024). *Report on the digitalisation of the European insurance sector*. https://www.eiopa.europa.eu/publications/eiopas-report-digitalisation-european-insurance-sector_en.
- European Network of National Human Rights Institutions. (2023). *ENNHRI common position on EU Artificial Intelligence Act*. ENNHRI. https://ennhri.org/wp-content/uploads/2023/09/ENNHRI-common-position-EU-AI-Act.pdf.
- European Parliament (2025) Press Release. MEPs debate proposed withdrawal of Horizontal Anti-Discrimination Directive.
- European Parliament and Council of the European Union. (2016). Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L 119, 1–88. https://eur-lex.europa.eu/eli/reg/2016/679/oj.European Parliament and Council of the European Union. (2022). Regulation (EU) 2022/2065 of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act). Official Journal of the European Union, L 277, 1–102. https://eur-lex.europa.eu/eli/reg/2022/2065/oj.
- European Parliament and Council Regulation (EU) 2026/1743 of 8 July 2026 amending Regulations (EU) 2016/679 and (EU) 2021/1232 as regards the simplification of certain obligations and requirements relating to data protection (Digital Omnibus on Data Protection) (Text with EEA relevance).
- European Parliament and of the Council Regulation (EU) 2026/1744 of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI) (Text with EEA relevance).
- European Parliament (2020), 2020/2014(INL) resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence.
- European Parliament and Council of the European Union. (2020). *Directive (EU) 2020/1828 of the European Parliament and of the Council of 25 November 2020 on representative actions for the protection of the collective interests of consumers and repealing Directive 2009/22/EC*. *Official Journal of the European Union, L 409*, 1–27. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32020L1828.
- European Union Agency for Fundamental Rights. (2014). *Violence against women:an EU wide survey: Main results.* Publications Office. https://doi.org/10.2811/62230.
- European Union Agency for Fundamental Rights, *Opinion on 'Improving access to remedy in the area of business and human rights at the EU level*, 2017, https://fra.europa.eu/en/opinion/2017/business-human-rights.
- European Union Agency for Fundamental Rights. (2020). *Handbook on European law relating to accessmto justice*.
- European Union Agency for Fundamental Rights. (2022). *Bias in algorithms: Artificial intelligence and discrimination*. Publications Office of the European Union. https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf.

- European Union Agency for Fundamental Rights (Ed.). (2023). *Online content moderation: Current challenges in detecting hate speech*. Publications Office. https://doi.org/10.2811/332335.
- European Union Agency for Fundamental Rights & European Institute for Gender Equality. (2026). *EU gender-based violence survey*. Publications Office of the European Union. https://doi.org/10.2811/1803934.
- Eurostat. (2026). *Women at work: A snapshot of EU's gender employment gap*. https://ec.europa.eu/eurostat/web/products-eurostat-news/w/edn-20260303-1.
- Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and Removing Disparate Impact. https://doi.org/10.48550/arXiv.1412.3756.
- Fernandez, M., Pérez, Á. P., Giallongo, D., Ghia, D., Yaqub, M., Quercia, D., & Cerquitelli, T. (2026). *When AI Enters the Workplace, Who Faces Greater Risks? A Gendered Analysis* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2609.21756.
- Fernández-Llorca, D., Gómez, E., Sánchez, I., & Mazzini, G. (2024). An interdisciplinary account of the terminological choices by EU policymakers ahead of the final agreement on the AI Act: AI system, general purpose AI system, foundation model, and generative AI. *Artificial Intelligence and Law*. https://doi.org/10.1007/s10506-024-09412-y.
- Ferraro, A., Ekstrand, M. D., & Bauer, C. (2024). It's Not You, It's Me: The Impact of Choice Models and Ranking Strategies on Gender Imbalance in Music Recommendation. *18th ACM Conference on Recommender Systems*, 884–889. https://doi.org/10.1145/3640457.3688163.
- Fioravanti, G., Bocci Benucci, S., Ceragioli, G., & Casale, S. (2022). How the Exposure to Beauty Ideals on Social Networking Sites Influences Body Image: A Systematic Review of Experimental Studies. *Adolescent Research Review*, *7*(3), 419–458. https://doi.org/10.1007/s40894-022-00179-4.
- Fosch-Villaronga, E., Mut-Piña, A., Verhoef, T., Poulsen, A., Søraa, R. A., Drukarch, H., Noel, A., Tunney, S., & Custers, B. (2025). User perceptions of misgendering algorithms. *Big Data & Society*, *12*(4), 20539517251398719. https://doi.org/10.1177/20539517251398719.
- Fosch-Villaronga, E., Poulsen, A., Søraa, R. A., & Custers, B. H. M. (2021). A little bird told me your gender: Gender inferences in social media. *Information Processing & Management*, *58*(3), 102541. https://doi.org/10.1016/j.ipm.2021.102541.
- Fox, S., & Epstein, D. A. (2020). Monitoring Menses: Design-Based Investigations of Menstrual Tracking Applications. In C. Bobel, I. T. Winkler, B. Fahs, K. A. Hasson, E. A. Kissling, & T.-A. Roberts (Eds), *The Palgrave Handbook of Critical Menstruation Studies* (pp. 733–750). Springer Singapore. https://doi.org/10.1007/978-981-15-0614-7_54.
- FRA & EIGE. (2026). *EU gender-based violence survey*. Publications Office of the European Union. https://doi.org/10.2811/6663401.
- Fredman, S. (2016). Substantive equality revisited. *International Journal of Constitutional Law*, 14(3), 712–738.
- Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, *14*(3), 330–347. https://doi.org/10.1145/230538.230561.

- Gallagher, R., & Jona, L. (2019, July 26). We tested Europe's new lie detector for travellers—And immediately triggered a false positive. *The Intercept*. https://theintercept.com/2019/07/26/europe-border-control-ai-lie-detector/.
- Gámez-Guadix, M., Almendros, C., Borrajo, E., & Calvete, E. (2015). Prevalence and Association of Sexting and Online Sexual Victimization Among Spanish Adults. *Sexuality Research and Social Policy*, *12*(2), 145–154. https://doi.org/10.1007/s13178-015-0186-9.
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H. M., III, H. D., & Crawford, K. (2018). *Datasheets for Datasets*. https://doi.org/10.48550/arXiv.1803.09010.
- Geena Davis Institute on Gender in Media. (2016). *The reel truth: Women aren't seen or heard*. https://geenadavisinstitute.org/research/the-reel-truth-women-arent-seen-or-heard/.
- Geiger, G., Constantaras, E., Braun, J.-C., Aung, H., Schot, E., Klassen, S., van Dijk, R., Davidson, D., Mehrota, D., Meaker, M., Burgess, M., Thomas, K., Walker, A., Lam, K., Lavigne, S., Qu, A., Nagendran, R., Moorthy, H., Tiwari, I., … Kollias, F. (2023). *Unprecedented experiment on welfare surveillance algorithm reveals discrimination*. Lighthouse Reports. https://www.lighthousereports.com/investigation/suspicion-machines/.
- Gerards, J. and Xenidis, R. (2021). *Algorithmic discrimination in Europe: Challenges and Opportunities for EU Gender Equality and Non-Discrimination Law*. European network of legal experts in gender equality and non-discrimination / European Commission, 2021.
- Gerards, J., Schäfer, M. T., Muis, I., & Vankan, A. (2022). *Fundamental Rights and Algorithms Impact Assessment (FRAIA)*. Government of the Netherlands. https://www.government.nl/documents/2021/07/31/impact-assessment-fundamental-rights-and-algorithms.
- Gerrard, Y., & Thornham, H. (2020). Content moderation: Social media's sexist assemblages. *New Media & Society*, *22*(7), 1266–1286. https://doi.org/10.1177/1461444820912540.
- The Gender Equality and Equal Treatment Commissioner's Office of Estonia, The Office of the Equal Opportunities Ombudsperson of Lithuania, Ministry of Economic Affairs and Communications of Estonia, Ministry of Justice of Estonia, & Tallinn University of Technology. (2024, November 25). *Legal analysis on non-discrimination regulation and regulatory gaps concerning algorithmic systems in Estonia and Lithuania*. https://www.volinik.ee/volinik-live-web-prd/s3fs-public/2026-02/legal_analysis_on_non-discrimination_regulation_and_regulatory_gaps_concerning_algorithmic_systems_in_estonia_and_lithuania_0.pdf.
- Gibson, C., Olszewski, D., Brigham, N. G., Crowder, A., Butler, K. R. B., Traynor, P., Redmiles, E. M., & Kohno, T. (2024). *Analyzing the AI Nudification Application Ecosystem* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2411.09751.
- Gill, R. (2021). Being watched and feeling judged on social media. *Feminist Media Studies*, *21*(8), 1387–1392. https://doi.org/10.1080/14680777.2021.1996427.
- Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
- Gkantona, G., Karakoula, P., & Kitsiou, A. (2026). Technology-Facilitated Violence Against Women in Greece: A Qualitative Study of Professionals' And Institutional Representatives' Responses and

Perspectives. *Journal of Aggression, Maltreatment & Trauma*, *35*(9), 1569–1594. https://doi.org/10.1080/10926771.2026.2718749.

- Glitch UK. (2023). *The Digital Misogynoir Report: Ending the Dehumanising of Black Women on Social Media*. https://glitchcharity.co.uk/wp-content/uploads/2023/07/Glitch-Misogynoir-Report_Final_18Jul_v5_Single-Pages.pdf.
- Global Witness. (2023). *New evidence of Facebook's sexist algorithm*. https://globalwitness.org/en/campaigns/digital-threats/new-evidence-of-facebooks-sexist-algorithm/.
- Glymour, B., & Herington, J. (2019). *Measuring the Biases that Matter: The Ethical and Casual Foundations for Measures of Fairness in Algorithms.* https://doi.org/10.1145/3287560.328757.
- Gómez Ortega, A., Van Der Maden, W., Lampinen, A., Comber, R., & Balaam, M. (2026). AI Loves Boobies: Unpacking Harmful Imaginaries in Community-Based Generative AI. *Proceedings of the 2026 Designing Interactive Systems Conference*, 158–171. https://doi.org/10.1145/3800645.3812786.
- Gonen, H., & Goldberg, Y. (2019). *Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them*. arXiv. https://arxiv.org/abs/1903.03862.
- Gourabathina, A., Gerych, W., Pan, E., & Ghassemi, M. (2025). The Medium is the Message: How Non-Clinical Information Shapes Clinical Decisions in LLMs. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 1805–1828. https://doi.org/10.1145/3715275.3732121.
- Government of Canada. (2025). *Algorithmic Impact Assessment (AIA)* (Version 1.0.1). Open Government Portal. https://open.canada.ca/aia-eia-js/?lang=en.
- Granberg, S., Geiger, G., Tiberg, A., Braun, J.-C., Malmsten, H., Molén, T., Abdigadir, A., Ljungmark, I., Laurin, F., Constantaras, E., & Howden, D. (2024). *Sweden's suspicion machine*. Lighthouse Reports and Svenska Dagbladet. https://www.lighthousereports.com/investigation/swedens-suspicion-machine/.
- Graux, H, Garstka, K., Murali, N., Cave, J., & Botterman, M. (2025). *Interplay between the AI Act and the EU digital legislative framework. Policy Department for Transformation, Innovation and Health, European Parliament.* https://www.europarl.europa.eu/RegData/etudes/STUD/2025/778575/ECTI_STU(2025)778575_EN.pdf.
- GREVIO. (2021). *General Recommendation No. 1 on the digital dimension of violence against women*. https://rm.coe.int/grevio-rec-no-on-digital-violence-against-women/1680a49147
- Griffin, R. (2025). *Governing platforms through corporate risk management: The politics of systemic risk in the Digital Services Act*. European Law Open, 4(2), 223–253. https://doi.org/10.1017/elo.2025.17.
- Grobman, W. A., Sandoval, G., Rice, M. M., Bailit, J. L., Chauhan, S. P., Costantine, M. M., Gyamfi-Bannerman, C., Metz, T. D., Parry, S., Rouse, D. J., Saade, G. R., Simhan, H. N., Thorp, J. M., Tita, A. T. N., Longo, M., & Landon, M. B. (2021). Prediction of vaginal birth after cesarean delivery in term gestations: A calculator without race and ethnicity. *American Journal of Obstetrics and Gynecology*, *225*(6), 664.e1–664.e7. https://doi.org/10.1016/j.ajog.2021.05.021.

- Guinea, O., Pandya, D., Sharma, V., & Zilli, R. (2025). *Mass litigation and the future of litigation funding in Ireland and Europe* (Occasional Paper No. 09/2025). European Centre for International Political Economy. https://ecipe.org/wp-content/uploads/2025/10/ECI_OccasionalPaper_09-2025_LY07.pdf.
- Hacker, P. (2018). Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review*, *55*(Issue 4), 1143–1185. https://doi.org/10.54648/COLA2018095.
- Hacker, P. (2023). *The European AI Liability Directives – Critique of a Half-Hearted Approach and Lessons for the Future*, 51 COMPUT. L. & SEC. REV., Nov. 2023, at 1, 28.
- Hague District Court, 5 February 2020, ECLI:NL:RBDHA:2020:865.
- Hall, L., & Clapton, W. (2021). Programming the machine: Gender, race, sexuality, AI, and the construction of credibility and deceit at the border. *Internet Policy Review*, *10*(4). https://doi.org/10.14763/2021.4.1601.
- HateAid, & Technische Universität München. (2025). *Angegriffen & alleingelassen: Wie sich digitale Gewalt auf politisches Engagement auswirkt. Ein Lagebild.* https://hateaid.org/wp-content/uploads/2025/01/hateaid-tum-studie-angegriffen-und-alleingelassen-2025.pdf.
- Hawkins, W., Mittelstadt, B., & Russell, C. (2025). Deepfakes on Demand: The rise of accessible non-consensual deepfake image generators. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 1602–1614. https://doi.org/10.1145/3715275.3732107.
- Hayes, C., Gupta, T., Crockett, C., & Bansal, V. (2026). *Engaging Men, Ending Violence: Pathways to Safer Digital Futures*. UN Women and Equimundo. https://www.unwomen.org/sites/default/files/2026-05/engaging-men-ending-violence-en.pdf.
- Hendl, T., & Jansky, B. (2022). Tales of self-empowerment through digital health technologies: A closer look at 'Femtech'. *Review of Social Economy*, *80*(1), 29–57. https://doi.org/10.1080/00346764.2021.2018027.
- Henry, N., & Umbach, R. (2024). Sextortion: Prevalence and correlates in 10 countries. *Computers in Human Behavior*, *158*, 108298. https://doi.org/10.1016/j.chb.2024.108298.
- Hill Collins, P., & Bilge, S. (2020). *Intersectionality* (2nd edition). Polity Press.
- Hoekstra, N. (2026). *Unlocking HER potential: Gender equality through equitable access to assistive technology*. UN Women. https://www.unwomen.org/en/digital-library/publications/2026/03/unlocking-her-potential-gender-equality-through-equitable-access-to-assistive-technology.
- Holland, S., Hosny, A., Newman, S., Joseph, J., Chmielinski, K. (2018). *The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards*. arXiv:1805.03677 [cs.DB].
- Home Security Heroes. (2023). *2023 State of deepfakes: Realities, threats, and impact*. https://www.securityhero.io/state-of-deepfakes/.
- Hooper, L., Remmits, Y., & Papageorgiou, I. (2026). *European assessment methodology on algorithmic discrimination*. Council of Europe. https://rm.coe.int/european-assessment-methodology-on-algorithmic-discrimination/48802dac82.

- Horwitz, J. (2025, August 29). *Meta created flirty chatbots of Taylor Swift, other celebrities without permission*. Reuters https://www.reuters.com/business/meta-created-flirty-chatbots-taylor-swift-other-celebrities-without-permission-2025-08-29/.
- Humlum, A., & Vestergaard, E. (2025). The unequal adoption of ChatGPT exacerbates existing inequalities among workers. *Proceedings of the National Academy of Sciences*, *122*(1), e2414972121. https://doi.org/10.1073/pnas.2414972121.
- Huyskes, D. (2026). 'Unprecedented Injustice': Network determinism in the Dutch digital welfare state. *AI and Ethics*, *6*(5), 527. https://doi.org/10.1007/s43681-026-01350-6.
- Illingworth, P. J., Venetis, C., Gardner, D. K., Nelson, S. M., Berntsen, J., Larman, M. G., Agresta, F., Ahitan, S., Ahlström, A., Cattrall, F., Cooke, S., Demmers, K., Gabrielsen, A., Hindkjær, J., Kelley, R. L., Knight, C., Lee, L., Lahoud, R., Mangat, M., … Hardarson, T. (2024). Deep learning versus manual morphology-based embryo selection in IVF: A randomized, double-blind noninferiority trial. *Nature Medicine*, *30*(11), 3114–3120. https://doi.org/10.1038/s41591-024-03166-5.
- ILO. (2026). *Gen AI, occupational segregation and gender equality in the world of work*. International Labour Organization. https://doi.org/10.54394/00033798.
- Jankowicz, N., Hunchak, J., Pavliuc, A., Davies, C., Pierson, S., & Kaufmann, Z. (2021). *Malign Creativity: How Gender, Sex, and Lies Are Weaponized against Women Online*. Wilson Center. https://www.wilsoncenter.org/sites/default/files/media/uploads/documents/Report%20Malign%20Creativity%20How%20Gender%2C%20Sex%2C%20and%20Lies%20are%20Weaponized%20Against%20Women%20Online_0.pdf.
- Jansen, H. A. F. M. (2023). *A guide to better understanding and using violence against women prevalence data*. UNFPA. https://www.unfpa.org/sites/default/files/pub-pdf/A-Guide-to-Better-Understanding%20%26%20Using-VAW-Prevalence-Data.pdf.
- Jiang, H., & Nachum, O. (2020). *Identifying and Correcting Label Bias in Machine Learning.* https://doi.org/10.48550/arXiv.1901.04966.
- Johnson, N., Moharana, S., Harrington, C. N., Andalibi, N., Heidari, H., & Eslami, M. (2024). *The fall of an algorithm: Characterizing the dynamics toward abandonment*. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. ACM. https://doi.org/10.1145/3630106.3658910.
- Judson, E. (2026). Gendered disinformation online targeting LGBT+ rights in Europe: What it looks like and how EU regulation can counter it. *Journal of Gender Studies*, *35*(3), 690–712. https://doi.org/10.1080/09589236.2026.2623941.
- Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1–33. https://doi.org/10.1007/s10115011-0463-8.
- Karagianni A. (2025). Gender in a stereo-(gender)typical EU AI law: A feminist reading of the AI act. Cambridge Forum on AI: Law and Governance. 2025;1:e25. https://doi.org/10.1017/cfl.2025.12.
- Karaian, L. (2024). Plastic fantastic: Sex robots and/as sexual fantasy. *Sexualities*, *27*(3), 633–652. https://doi.org/10.1177/13634607221106667.

- Kaun, A., & Liminga, A. (2025). Welfare service centers: Maintenance, repair, and care at the analog interfaces of the digital welfare state. *New Media & Society*, *27*(5), 3039–3054. https://doi.org/10.1177/14614448231220362.
- Kayser-Bril, N., Richard, É., Duportail, J., & Schacht, K. (2020). *Undress or fail: Instagram's algorithm strong-arms users into showing skin*. AlgorithmWatch and the European Data Journalism Network. https://algorithmwatch.org/en/instagram-algorithm-nudity/.
- Keller, P. & Aplin, T. (2026). *Reconciling trade secrets and AI public transparency*. *The Cambridge Law Journal, 85*(1), 151-185. https://doi.org/10.1017/S0008197325101086.
- Kelley, S., Ovchinnikov, A., Hardoon, D. R., & Heinrich, A. (2022). Antidiscrimination Laws, Artificial Intelligence, and Gender Bias: A Case Study in Nonmortgage Fintech Lending. *Manufacturing & Service Operations Management*, *24*(6), 3039–3059. https://doi.org/10.1287/msom.2022.1108.
- Kelly, L. (1988). *Surviving sexual violence*. Polity Press; B. Blackwell.
- Kelly, T. (2022). *Bias: A Philosophical Study* (1st edn). Oxford University Press. https://doi.org/10.1093/oso/9780192842954.001.0001.
- Kettemann, M.C. and Tiedeke, A.S., 2020. *Back up: Can users sue platforms to reinstate deleted content?*. Internet Policy Review, 9(2), pp.1-20.
- Keyes, O. (2018). The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–22. https://doi.org/10.1145/3274357.
- Knecht, A., & Pühringer, J. (2018, November 2). Eine Formel für Gerechtigkeit Der AMS-Algorithmus verdeutlicht vorhandene Diskriminierungen am Arbeitsmarkt. *Der Standard*. https://www.derstandard.at/2000090547505/eine-formel-fuer-gerechtigkeit.
- Krasanakis, E., Xioufis, E. S., Papadopoulos, S., & Kompatsiaris, Y. (2018). *Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-aware Classification.* Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-aware Classification | Proceedings of the 2018 World Wide Web Conference.
- Koukopoulos, N., Janickyj, M., & Tanczer, L. M. (2026). Defining and Conceptualizing Technology-Facilitated Abuse ('Tech Abuse'): Findings of a Global Delphi Study. *Journal of Interpersonal Violence*, *41*(1–2), 249–275. https://doi.org/10.1177/08862605241310465.
- Köver, C., & Schwarzbeck, M. (2025, January 27). *mSpy-Leak: Vom Partner verwanzt* [mSpy leak: Bugged by one's partner]. netzpolitik.org. https://netzpolitik.org/2025/mspy-leak-vom-partner-verwanzt/.
- Köver, C., & Schwarzbeck, M. (2026, February 13). *Digitales Stalking: „Er wusste immer genau, wo ich war"*. netzpolitik.org. https://netzpolitik.org/2026/digitales-stalking-er-wusste-immer-genau-wo-ich-war/.
- KU Leuven, Medical University of Plovdiv, Maggioli Spa, Democracy X, Institute of Philosophy and Technology, Trustilio, Valencia Port, IBM (France), & 21c Consultancy (United Kingdom). (2026). Full Policy Brief: Socio-Political, Legal & Ethical Recommendations (for building trustworthy AI). Zenodo. https://doi.org/10.5281/zenodo.22933991.
- Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). *Counterfactual Fairness* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1703.06856.

- Kužel, R., Adamčíková, J., Mračka, M., & Godársky, I. (2025). *Gendered disinformation in Slovakia & Czechia*. MEMO 98; Ženy v médiích. https://www.memo98.sk/storage/media/2026/05/13/32724236-1d6e-4051-bc08-dbe1cd377925.pdf.
- Kwarsick, J. (2026), Why AI Adoption Fails: A Constraint-Based Diagnostic Model for Organizational Readiness, Available at SSRN: https://ssrn.com/abstract=6446445 or http://dx.doi.org/10.2139/ssrn.6446445.
- Lambrecht, A., & Tucker, C. (2019). Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science*, *65*(7), 2966–2981. https://doi.org/10.1287/mnsc.2018.3093.
- Lång, K., Josefsson, V., Larsson, A.-M., Larsson, S., Högberg, C., Sartor, H., Hofvind, S., Andersson, I., & Rosso, A. (2023). Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): A clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *The Lancet Oncology*, *24*(8), 936–944. https://doi.org/10.1016/S1470-2045(23)00298-X.
- Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., & Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, *117*(23), 12592–12594. https://doi.org/10.1073/pnas.1919012117.
- Lau, P. L. (2024). 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* (Vol. 2023, pp. 247–274). Springer Nature Switzerland. https://doi.org/10.1007/16495_2023_63.
- Laux, J., & Ruschemeier, H. (2025). Automation Bias in the AI Act: On the Legal Implications of Attempting to De-Bias Human Oversight of AI. *European Journal of Risk Regulation*, *16*(4), 1519–1534. https://doi.org/10.1017/err.2025.10033.
- Lázaro Cabrera, L., & Maier, M. (2026). *A long road ahead: Effective remedies for artificial intelligence harms in the European Union*. Centre for Democracy and Technology Europe. https://cdt.org/wp-content/uploads/2026/03/CDT-Remedies-for-AI-Harms.pdf.
- Leerssen, P., van Duin, A., Toepoel, I., & van Hoboken, J. (2025), *Pathways to private enforcement of the Digital Services Act (DSA)*. Institute for Information Law (IViR), University of Amsterdam. https://dsa-observatory.eu/wp-content/uploads/2025/06/DSA-Private-Enforcement-final-draft.pdf.
- Lenaerts (2020). *The Principle of Effective Judicial Protection in EU Law*; Prechal, S. (2015). *The Court of Justice and Effective Judicial Protection: What Has the Charter Changed?*
- Leupold, P. (2022, August). *Private international law and cross-border collective redress: A legal analysis of jurisdiction, applicable law, pendency, recognition and enforcement under the Representative Actions Directive 1828/2020* (BEUC-X-2022-085). BEUC. https://www.beuc.eu/sites/default/files/publications/BEUC-X-2022-085_Private_International_Law_and_Cross-Border_Collective_Redress.pdf.
- Lewis, R., & Anitha, S. (2022). Upskirting: A Systematic Literature Review. *Trauma, Violence, & Abuse*, 152483802210820. https://doi.org/10.1177/15248380221082091.

- Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2022). Discrimination-free insurance pricing. *ASTIN Bulletin*, *52*(1), 55–89. https://doi.org/10.1017/asb.2021.23.
- Liu, M., Ke, Y., Zhu, W., Mertens, M., Ning, Y., Liao, J., Hong, C., Ting, D. S. W., Peng, Y., Bitterman, D. S., Ong, M. E. H., & Liu, N. (2025). *Gender Bias in Large Language Models for Healthcare: Assignment Consistency and Clinical Implications* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2510.08614.
- Lomba, N., Navarra, C., & Fernandes, M. (2021). *Combating gender-based violence: Cyberviolence—European added value assessment*. https://www.europarl.europa.eu/RegData/etudes/STUD/2021/662621/EPRS_STU(2021)662621_EN.pdf.
- Lorenz, P., Perset, K., & Berryhill, J. (2023). *Initial policy considerations for generative artificial intelligence*. OECD, https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/09/initial-policy-considerations-for-generative-artificial-intelligence_1a9ab450/fae2d1e6-en.pdf.
- Lulli, R., Sinicato, F., & Couceiro Farjas, Á. A. (2024). *Digital Skills, Accommodation and Technological Assistance for Employment: Supporting the inclusion of persons with disabilities in the open labour market*. European Disability Forum, https://www.edf-feph.org/content/uploads/2024/10/DATA-project-summary-report.pdf.
- Lungarella, M., Iida, F., Bongard, J. C., & Pfeifer, R. (2007). AI in the 21st Century – With Historical Reflections. In M. Lungarella, F. Iida, J. Bongard, & R. Pfeifer (Eds), *50 Years of Artificial Intelligence* (Vol. 4850, pp. 1–8). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-77296-5_1.
- Malgieri & Fuster, (2022), t*he European Data Protection Board (EDPB) and the European Data Protection Supervisor (EDPS) recommended a ban on AI biometric categorisation systems that classify individuals based on gender*.
- Mantelero, A. (2024). The Fundamental Rights Impact Assessment (FRIA) in the AI Act: Roots, legal obligations and key elements for a model template. *Computer Law & Security Review, 54*, 106020. https://doi.org/10.1016/j.clsr.2024.106020.
- Martinková, S., Stańczak, K., & Augenstein, I. (2023). *Measuring Gender Bias in West Slavic Language Models* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2304.05783.
- McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence: August 31, 1955*. 12–14.
- McCorduck, P. (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence* (25th anniversary update). A.K. Peters.
- McGlynn, C. (2024). Towards a New Criminal Offence of Intimate Intrusions. *Feminist Legal Studies*. https://doi.org/10.1007/s10691-024-09547-y.
- McGlynn, C., & Johnson, K. (2021). *Cyberflashing: Recognising Harms, Reforming Laws* (1st edn). Bristol University Press. https://doi.org/10.2307/j.ctv1jk0hd4.
- McGlynn, C., Johnson, K., Rackley, E., Henry, N., Gavey, N., Flynn, A., & Powell, A. (2021). 'It's Torture for the Soul': The Harms of Image-Based Sexual Abuse. *Social & Legal Studies*, *30*(4), 541–562. https://doi.org/10.1177/0964663920947791.

- McGlynn, C., McDermott, Y., Macdonald, S., Toparlak, R. T., Tarrant, F., & Treacy, S. (2026). *AI Chatbots and Violence Against Women and Girls: New Frontiers, New Harms*. UK Research and Innovation. https://e87dab74-be98-4bb1-83c5-05251d2bc6f4.usrfiles.com/ugd/e87dab_06a7f0801de549689c294d42e0478a3c.pdf.
- McGlynn, C., & Rackley, E. (2017). Image-Based Sexual Abuse. *Oxford Journal of Legal Studies*, *37*(3), 534–561. https://doi.org/10.1093/ojls/gqw033.
- McGlynn, C., Rackley, E., & Houghton, R. (2017). Beyond 'Revenge Porn': The Continuum of Image-Based Sexual Abuse. *Feminist Legal Studies*, *25*(1), 25–46. https://doi.org/10.1007/s10691-017-9343-2.
- McGlynn, C., & Rigotti, C. (2025). From Virtual Rape to Meta-rape: Sexual Violence, Criminal Law and the Metaverse. *Oxford Journal of Legal Studies*, gqaf009. https://doi.org/10.1093/ojls/gqaf009.
- McGlynn, C., & Toparlak, R. T. (2025). The 'new voyeurism': Criminalizing the creation of 'deepfake porn'. *Journal of Law and Society*, *52*(2), 204–228. https://doi.org/10.1111/jols.12527.
- McGregor, L., Murray, D., & Ng, V. (2019). *International human rights law as a framework for algorithmic accountability*. *International & Comparative Law Quarterly, 68*(2), 309-343. https://doi.org/10.1017/S0020589319000046.
- McMillan, C. (2024). Gender, health and technology: The rise of femtech – periods, profit and power. In J. Gideon & S. Hawkes (Eds), *A Research Agenda for Gender and Health* (pp. 87–106). Edward Elgar Publishing. https://doi.org/10.4337/9781802209228.00010.
- Meding, K. (2025). *It's complicated. The relationship of algorithmic fairness and non-discrimination regulations for high-risk systems in the EU AI Act* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2501.12962.
- Merone, L., Tsey, K., Russell, D., & Nagle, C. (2022). Sex Inequalities in Medical Research: A Systematic Scoping Review of the Literature. *Women's Health Reports*, *3*(1), 49–59. https://doi.org/10.1089/whr.2021.0083.
- Ministerio de la Presidencia, Justicia y Relaciones con las Cortes. (2026a). *Resolución de 10 de abril de 2026, de la Subsecretaría, por la que se publica el Convenio entre la Agencia Española de Supervisión de Inteligencia Artificial y el Instituto de las Mujeres, O.A., para la colaboración en la supervisión de sistemas de IA, la protección de los derechos fundamentales, la alfabetización y el desarrollo de laboratorios de ideas en materia de Inteligencia Artificial* (BOE-A-2026-8672). *Boletín Oficial del Estado*, No. 96. https://www.boe.es/diario_boe/txt.php?id=BOE-A-2026-8672.
- Ministerio de la Presidencia, Justicia y Relaciones con las Cortes. (2026b). *Real Decreto 634/2026, de 29 de julio, por el que se aprueban los Estatutos del Instituto de las Mujeres, O.A.* (BOE-A-2026-16550). *Boletín Oficial del Estado*, No. 185. https://www.boe.es/buscar/doc.php?id=BOE-A-2026-16550.
- Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, *8*(1), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902.

- Mittelstadt, B., Wachter, S. & Russell, C. (2024). *The Unfairness of Fair Machine Learning: Leveling Down and Strict Egalitarianism by Default*, 30 Mich. Tech. L. Rev. (2024). Available at: https://repository.law.umich.edu/mtlr/vol30/iss1/3.
- Moldovan, A., Csürös, K., Bucur, A., & Bercuci, L. (2022). Users hate blondes: Detecting sexism in user comments on online Romanian news. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.woah-1.21.
- Montgomery, B. (2024, January 31). Taylor Swift AI images prompt US bill to tackle nonconsensual, sexual deepfakes. *The Guardian*. https://www.theguardian.com/technology/2024/jan/30/taylor-swift-ai-deepfake-nonconsensual-sexual-images-bill.
- Morgan, S., Farid, H., & Nightingale, S. J. (2026). *The backbone of abuse: How infrastructure providers enable the proliferation of AI-generated non-consensual intimate imagery*. *Journal of Online Trust and Safety, 3*(4). https://doi.org/10.54501/jots.v3i4.348.
- Narayanan, A., & Kapoor, S. (2024). *AI snake oil: What artificial intelligence can do, what it can't, and how to tell the difference*. Princeton University Press.
- Neff, G., & Nagy, P. (2016). Automation, Algorithms, and Politics| Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal of Communication*, *10*, 4915–4931.
- Ng, A. Y., Oberije, C. J. G., Ambrózay, É., Szabó, E., Serfőző, O., Karpati, E., Fox, G., Glocker, B., Morris, E. A., Forrai, G., & Kecskemethy, P. D. (2023). Prospective implementation of AI-assisted screen reading to improve early detection of breast cancer. *Nature Medicine*, *29*(12), 3044–3049. https://doi.org/10.1038/s41591-023-02625-9.
- Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics, 2*(1), 25-42. https://doi.org/10.1007/BF02639315.
- NL Times. (2023, October 19). *Man faces community service for making deepfake porn of TV host Welmoed Sijtsma*. https://nltimes.nl/2023/10/19/man-faces-community-service-making-deepfake-porn-tv-host-welmoed-sijtsma.
- NOS. (2024b, March 19). *Vrouwelijke BN'ers overwegen aangifte vanwege deepfake pornovideo's* [Female Dutch public figures consider reporting deepfake porn videos]. NOS Nieuws. https://nos.nl/l/2513371.
- noyb – European Center for Digital Rights. (2026). Digital Omnibus: First analysis of select GDPR and ePrivacy proposals by the Commission (Version 3.0). https://noyb.eu/sites/default/files/2025-12/noyb%20Digital%20Omnibus%20Report%20V1.pdf.
- OAS, & MESECVI. (2025). *Inter-American Model Law to Prevent, Punish, and Eradicate Gender-Based Digital Violence Against Women*. https://belemdopara.org/wp-content/uploads/2026/03/EN-Ley-Modelo_MESECVI-CEVI-1.pdf.
- OECD. (2023). *The Impact of AI on the Workplace: Evidence from OECD Case Studies of AI Implementation* (OECD Social, Employment and Migration Working Papers No. 289; OECD Social, Employment and Migration Working Papers, Vol. 289). https://doi.org/10.1787/2247ce58-en.
- OECD. (2024). *Algorithm and Eve: How AI will impact women at work*. https://www.oecd.org/en/publications/algorithm-and-eve_a1603510-en.html.

- OHCHR. (2018). *Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective*.
- O'Malley, R. L. (2023). Short-Term and Long-Term Impacts of Financial Sextortion on Victim's Mental Well-Being. *Journal of Interpersonal Violence*, *38*(13–14), 8563–8592. https://doi.org/10.1177/08862605231156416.
- O'Malley, R. L., Bolton, M., Ayerza, J., & Colavito, K. (2026). 'I Don't Feel Safe at All': Examining Shattered Assumptions Among Male Victims of Sextortion. *Journal of Interpersonal Violence*, 08862605261446971. https://doi.org/10.1177/08862605261446971.
- O'Malley, R. L., & Holt, K. M. (2022). Cyber Sextortion: An Exploratory Analysis of Different Perpetrators Engaging in a Similar Crime. *Journal of Interpersonal Violence*, *37*(1–2), 258–283. https://doi.org/10.1177/0886260520909186.
- Otokiti, A. U., Shih, H., & Williams, K. S. (2025). Gender and racial bias unveiled: Clinical artificial intelligence (AI) and machine learning (ML) algorithms are fanning the flames of inequity. *Oxford Open Digital Health*, *3*, oqaf027. https://doi.org/10.1093/oodh/oqaf027.
- Oudshoorn, N., Rommes, E., & Stienstra, M. (2004). Configuring the User as Everybody: Gender and Design Cultures in Information and Communication Technologies. *Science, Technology, & Human Values*, *29*(1), 30–63. https://doi.org/10.1177/0162243903259190.
- Opuz v. Turkey, App. No. 33401/02, Eur. Ct. H.R. (2009). https://hudoc.echr.coe.int/eng?i=001-92945.
- Palęcka, A., & Sztandar-Sztanderska, K. (2024). Aktywizacja czy dezaktywizacja zawodowa matek małych dzieci w powiatowych urzędach pracy? Upłciowienie norm dotyczących pracy. *Studia Socjologiczne*, 117–138. https://doi.org/10.24425/sts.2024.151936.
- Paliokaitė, A., Christenko, A., Aloisi, A., Moore, P., Potocka-Sionek, N., Donoghue, R. S., Krūminas, P., Čop, D., Bednorz, J., Daminov, I., & Ragaliauskaitė, A. (2025). *Study exploring the context, challenges, opportunities, and trends in algorithmic management in the workplace: Final report.* European Commission. https://data.europa.eu/doi/10.2767/5629841.
- Paradiso, M. N., Rollè, L., & Trombetta, T. (2023). Image-Based Sexual Abuse Associated Factors: A Systematic Review. *Journal of Family Violence*. https://doi.org/10.1007/s10896-023-00557-z.
- Park, K., Ging, D., Murphy, S., & McGrath, C. (2023). *The impact of the use of social media on women and girls*. European Parliament. https://www.europarl.europa.eu/RegData/etudes/STUD/2023/743341/IPOL_STU(2023)743341_EN.pdf.
- Passi, S., & Barocas, S. (2019). Problem Formulation and Fairness. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 39–48. https://doi.org/10.1145/3287560.3287567.
- Patel, N. J. (2021, December 21). Reality of Fiction? Sexual Harassment in VR, The Proteus Effect and the phenomenology of Darth Vader—And other stories.... *Medium*. https://medium.com/kabuni/fiction-vs-non-fiction-98aa0098f3b0.
- Posetti, J., Hellmueller, L., Williams, K., Renaud, P., Aboulez, N., & Shabbir, N. (2025). *Tipping point: The chilling escalation of online violence against women in the public sphere*. UN Women. https://www.unwomen.org/sites/default/files/2025-12/tipping-point-the-chilling-escalation-of-violence-against-women-in-the-public-sphere-in-the-age-of-ai-en.pdf.

- Powell, A. (2022). Technology-Facilitated Sexual Violence. In M. A. H. Horvath & J. M. Brown, *Rape* (2nd edn, pp. 143–158). Routledge. https://doi.org/10.4324/9781003163800-13.
- Puttick, A., Ikae, C., Rigotti, C., Fosch-Villaronga, E., Kharas, M. W., Søraa, R. A., & Kurpicz-Briki, M. (2025). A systematic review of bias detection methods for non-English word embeddings and language models. *Artificial Intelligence Review*, *58*(12), 370. https://doi.org/10.1007/s10462-025-11375-8.
- Qiwei, L., Santo, W. L., Schoenebeck, S., & Gilbert, E. (2026). *Position: AI/ML Deepfake Research is Misaligned with AI-Generated Non-Consensual Intimate Imagery (AIG-NCII)* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2607.18263.
- Rafnsdóttir, G. L., Marrero, A. S., Rigotti, C., Haraldsdóttir, R. K., & Fosch-Villaronga, E. (2026). Artificial Intelligence-Driven Inequality: Gendered and Intersectional Experiences of Workplace Discrimination in Europe. *Gender, Work & Organization*, gwao.70239. https://doi.org/10.1111/gwao.70239.
- Ray, A., & Henry, N. (2025). Sextortion: A Scoping Review. *Trauma, Violence, & Abuse*, *26*(1), 138–155. https://doi.org/10.1177/15248380241277271.
- Reuters. (2025, August 29). *Italy's Meloni "disgusted" by websites targeting women amid outcry over online abuse*. https://www.reuters.com/sustainability/society-equity/italys-meloni-disgusted-by-websites-targeting-women-amid-outcry-over-online-2025-08-29/.
- Riccio, P., Colin, J., Ogolla, S., & Oliver, N. (2024). Mirror, Mirror on the Wall, Who Is the *Whitest* of All? Racial Biases in Social Media Beauty Filters. *Social Media + Society*, *10*(2), 20563051241239295. https://doi.org/10.1177/20563051241239295.
- Richtlijn 2000/78/EG van de Raad van 27 november 2000 tot instelling van een algemeen kader voor gelijke behandeling in arbeid en beroep, CONSIL, 303 OJ L (2000). http://data.europa.eu/eli/dir/2000/78/oj.
- Rickman, S. (2025). Evaluating gender bias in large language models in long-term care. *BMC Medical Informatics and Decision Making*, *25*(1), 274. https://doi.org/10.1186/s12911-025-03118-0.
- Rigotti, C. (2025). *The regulation of sex robots: Gender and sexuality in the era of artificial intelligence*. Routledge.
- Rigotti, C., & Fosch-Villaronga, E. (2024). Fairness, AI & recruitment. *Computer Law & Security Review*, *53*, 105966. https://doi.org/10.1016/j.clsr.2024.105966.
- Rigotti, C., & McGlynn, C. (2022). Towards a European criminal law on violence against women: The ambitions and limitations of the Commission proposal to criminalise image-based sexual abuse. *New Journal of European Criminal Law*, *13*(4), 452–477.
- Rigotti, C., & McGlynn, C. (2025). Online and technology-facilitated violence against women: The EDVAW Platform's contribution to human rights protection and monitoring. *Netherlands Quarterly of Human Rights*, 82–103. https://doi.org/10.1177/09240519251334080.
- Rigotti, C., Pasvanti Gkioka, C., Do Rosario Da Graça, S., & Lannier, S. (2025). Overlapping offences, divergent logics: Reconciling EU law on sexual imagery of minors. New Journal of European Criminal Law, 20322844251401415. https://doi.org/10.1177/20322844251401415.

- Rigotti, C. (2026a). Sexual content, platform governance, and the Digital Services Act. *International Journal of Law and Information Technology, 34*, eaag017. https://doi.org/10.1093/ijlit/eaag017.
- Rigotti, C. (2026b, forthcoming). A necessary but imperfect fix: Article 5(1)(ba) and (bb) on non-consensual intimate and child sexual abuse material. In G. Malgieri, G. González Fuster, A. Mantelero, & G. Zanfir-Fortuna (Eds.), The EU Artificial Intelligence Act: A thematic commentary. Hart Publishing.
- Rigotti, C., Pasvanti Gkioka, C., Do Rosario Da Graça, S., & Lannier, S. (2025). Overlapping offences, divergent logics: Reconciling EU law on sexual imagery of minors. *New Journal of European Criminal Law*, 20322844251401415. https://doi.org/10.1177/20322844251401415.
- Ringrose, J., Milne, B., Mishna, F., Regehr, K., & Slane, A. (2022). Young people's experiences of image-based sexual harassment and abuse in England and Canada: Toward a feminist framing of technologically facilitated sexual violence. *Women's Studies International Forum*, *93*, 102615. https://doi.org/10.1016/j.wsif.2022.102615.
- Ritterbusch, G. D., & Teichmann, M. R. (2023). Defining the Metaverse: A Systematic Literature Review. *IEEE Access*, *11*, 12368–12377. https://doi.org/10.1109/ACCESS.2023.3241809.
- Roth-Cohen, O., Kanevska, H. S., & Eisend, M. (2023). Gender roles in online advertising. *Journal of Gender Studies*, *32*(2), 186–200. https://doi.org/10.1080/09589236.2022.2102970.
- Rousaki, A., & Fido, D. (2026). Understanding Image-Based Sexual Abuse Through Greek Public and Legal Lenses: A Constructionist Thematic Analysis. *Journal of Interpersonal Violence*, 08862605251411875. https://doi.org/10.1177/08862605251411875.
- Rudinger, R., Naradowsky, J., Leonard, B., & Van Durme, B. (2018). *Gender Bias in Coreference Resolution* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1804.09301.
- Ruschemeier, H. (2023). AI as a challenge for legal regulation – the scope of application of the artificial intelligence act proposal. *ERA Forum*, *23*(3), 361–376. https://doi.org/10.1007/s12027-022-00725-6.
- Russell, S. and Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson.
- Russell, S., Norvig, P., and Davis, E. (2010). Artificial intelligence: A modern approach (3rd ed). Prentice Hall.
- Salimi, B., Rodriguez, L., Howe, B., & Suciu, D. (2019). Interventional Fairness: Causal Database Repair for Algorithmic Fairness. Interventional Fairness | Proceedings of the 2019 International Conference on Management of Data.
- Sandu, G., & Cséby, R. B. (2026). *Women's rights and democracy: Combatting stereotypes, disinformation, violence in the digital age* (Briefing No. PE 785.000). European Parliament, Policy Department for Citizens, Equality and Culture.
- Saner, K., Barker, K., Gehrke, M., Łabuz, M., Stolze, M., Hooper, L., & Sessa, M. G. (2026). Gender and identity disinformation: Impacts on female politicians, and implications for democratic participation – roundtable discussion. *Journal of Gender Studies*, *35*(3), 713–729. https://doi.org/10.1080/09589236.2026.2637527.
- Satariano, A., & Toll Pifarré, R. (2024, July 18). An algorithm told police she was safe. Then her husband killed her. *The New York Times*.

https://www.nytimes.com/interactive/2024/07/18/technology/spain-domestic-violence-viogen-algorithm.html.

- Sattigeri, P., Hoffman, S. C., Chenthamarakshan, V., & Varshney, K. R. (2019). *Fairness GAN*. https://doi.org/10.48550/arXiv.1805.09910.
- Semenzin, S., & Bainotti, L. (2020). The use of Telegram for non-consensual dissemination of intimate images: Gendered affordances and the construction of masculinities. *Social Media + Society*, *6*(4). https://doi.org/10.1177/2056305120984453.
- Scally, D. (2026, April 1). How AI porn allegations surrounding TV star rocked Germany. *The Irish Times*. https://www.irishtimes.com/world/europe/2026/04/01/how-ai-porn-allegations-surrounding-tv-star-rocked-germany/.
- Schellevis, J., & van Hulzen, D. (2023, September 2). *10.000 berichten per dag in Nederlandse wraakpornogroepen: "Telegram doet niks"*. NOS Nieuws. https://nos.nl/l/2488889.
- Schoenebeck, S., Lampe, C., & Triệu, P. (2023). Online Harassment: Assessing Harms and Remedies. *Social Media + Society*, *9*(1), 20563051231157297. https://doi.org/10.1177/20563051231157297.
- Scholten, M., Arroyo Jiménez, L., & Martín Delgado, I. (2024). *Artificial Intelligence and EU law Enforcement: A Bottom-Up Approach. European Papers - A Journal on Law and Integration*, *9*(3), 930-934. https://doi.org/10.15166/2499-8249/792.
- Schulenberg, K., Freeman, G., Li, L., & Barwulor, C. (2023). "Creepy towards my avatar body, creepy towards my body": How women experience and manage harassment risks in social virtual reality. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2), Article 236. https://doi.org/10.1145/3610027.
- Schwarcz, D., & Prince, A. E. R. (2020). Proxy Discrimination in the Age of Artifificial Intelligence and Big Data. *Iowa Law Review*, *105*, 1257–1318.
- Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence, National Institute of Standards and Technology (U.S.)*. https://doi.org/10.6028/NIST.SP.1270.
- Sekwenz, M.-T., Beadle, K., & Parkin, S. (2026).*'There is literally zero funding': Understanding the emerging role of trusted flaggers under the EU Digital Services Act* (arXiv:2603.29874). *arXiv*. https://doi.org/10.48550/arXiv.2603.29874.
- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. https://doi.org/10.1145/3287560.3287598.
- Semenzin, S., & Bainotti, L. (2020). The Use of Telegram for Non-Consensual Dissemination of Intimate Images: Gendered Affordances and the Construction of Masculinities. *Social Media + Society*, *6*(4), 205630512098445. https://doi.org/10.1177/2056305120984453.
- Semenzin, S., Kerby, N., Schüler, M., & Romano, S. (2026). *Harassment As Infrastructure: How Telegram's design enables technology-facilitated gender-based violence in Italy and Spain*. AI Forensics. https://aiforensics.org/uploads/NCII%20Telegram%20-%20Harassment%20As%20Infrastructure.pdf.
- Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-

served patient populations. *Nature Medicine*, *27*(12), 2176–2182. https://doi.org/10.1038/s41591-021-01595-0.

- Shah, A. S. V., Griffiths, M., Lee, K. K., McAllister, D. A., Hunter, A. L., Ferry, A. V., Cruikshank, A., Reid, A., Stoddart, M., Strachan, F., Walker, S., Collinson, P. O., Apple, F. S., Gray, A. J., Fox, K. A. A., Newby, D. E., & Mills, N. L. (2015). High sensitivity cardiac troponin and the under-diagnosis of myocardial infarction in women: Prospective cohort study. *BMJ*, g7873. https://doi.org/10.1136/bmj.g7873.
- Short, E., McNamara, N., Pitchford, M., Conradie, L., Myers, E., Bhaker, S., & Barnes, J. (2026). 'What I Would Have Liked from the Police': Victims' Experiences of Police Engagement During the Investigation of Cyberstalking and Cyberharassment. *Journal of Victimology and Victim Justice*, 25166069261435357. https://doi.org/10.1177/25166069261435357.
- Shrishak, K., & Pénicaud, S. (2025). *European policy guidelines on AI and algorithm-driven discrimination for equality bodies and other national human rights structures*. Council of Europe.
- Shrishak, K. (2024). *D1 – AI bias evaluation. European Data Protection Board. https://www.edpb.europa.eu/documents/support-pool-of-experts/ai-complex-algorithms-and-effective-data-protection-supervision_en.*
- Simon, P.-A., Denuit, M., & Trufin, J. (2026). DPTree and DPForest: Tree-based methods fulfilling demographic parity. *Annals of Actuarial Science*, *20*(1), 96–114. https://doi.org/10.1017/S1748499525100092.
- Sivakaminathan, S. S., & Musi, E. (2026). ChatGPT is a gender bias echo-chamber in HR recruitment: An NLP analysis and framework to uncover the language roots of bias. *AI & SOCIETY*, *41*(4), 2841–2861. https://doi.org/10.1007/s00146-025-02564-8.
- Sjöholm, M. (2024). Regulation of online gender-based hate speech and international human rights law: Current status and challenges. *QIL, Zoom-In*, *104*, 3–16.
- Smuha, N. (2021). *Beyond the Individual: Governing AI's Societal Harm*. Internet Policy Review, 10(3). https://doi.org/10.14763/2021.3.1574.
- Smuha, N. (2025). *Algorithmic rule by law: How algorithmic regulation in the public sector erodes the rule of law*. Cambridge University Press. https://doi.org/10.1017/9781009427500.
- Smuha, N. & Yeung, K (2025). *The European Union's AI Act: Beyond motherhood and apple pie?* SSRN. https://doi.org/10.2139/ssrn.4874852. in N.A. Smuha (ed.) *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence*. Cambridge: Cambridge University Press (Cambridge Law Handbooks), pp. 228–258.
- Soler Garrido, J., De Nigris, S., Bassani, E., Sanchez, I., Tatjana, E., Antoine-Alexandre, A., & Thierry, B. (2024). *Harmonised Standards for the European AI Act*. European Commission. https://publications.jrc.ec.europa.eu/repository/handle/JRC139430.
- Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019). Evaluating Gender Bias in Machine Translation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1679–1684. https://doi.org/10.18653/v1/P19-1164.
- Steckelmacher, D., & Bogaerts, B. (2025). Definition of AI. In V. L. Raposo (Ed.), *The European Artificial Intelligence Act* (Vol. 78, pp. 81–103). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-98406-8_4.

- Stilinovic, M., Egliston, B., Tan, W., Carter, M., & Gray, J. E. (2026). *Harm Through the '#RizzCam': Smart Glasses, Ambient Capture and Invisible Harassment*. SocArXiv. https://doi.org/10.31235/osf.io/wbj2x_v2.
- Straw, I., & Wu, H. (2022). Investigating for bias in healthcare algorithms: A sex-stratified analysis of supervised machine learning models in liver disease prediction. *BMJ Health & Care Informatics*, *29*(1), e100457. https://doi.org/10.1136/bmjhci-2021-100457.
- Strikwerda, L. (2015). Present and Future Instances of Virtual Rape in Light of Three Categories of Legal Philosophical Theories on Rape. *Philosophy & Technology*, *28*(4), 491–510. https://doi.org/10.1007/s13347-014-0167-6.
- Sun, Y., Parker, T., Gierschmann, L., Shams, S., Canmetin, T., Duteil, M., Gipiškis, R., & Chin, Z. S. (2026). *Defining AI Models and AI Systems: A Framework to Resolve the Boundary Problem* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2603.10023.
- Suresh, H., & Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9. https://doi.org/10.1145/3465416.3483305.
- Swedish Medical Products Agency. (2018). *The assessment of Natural Cycles is completed*. https://via.tt.se/pressmeddelande/3249375/the-assessment-of-natural-cycles-is-completed?publisherId=3235477.
- Tatman, R. (2017). Gender and Dialect Bias in YouTube's Automatic Captions. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 53–59. https://doi.org/10.18653/v1/W17-1606.
- Thakur, D., & Hankerson, D. L. (2022). *Facts and their Discontents: A Research Agenda for Online Disinformation, Race, and Gender*. Open Science Framework. https://doi.org/10.31219/osf.io/3e8s5.
- Tribunale di Bologna, Sezione Lavoro. (2020, December 31). *Ordinanza, R.G. n. 2949/2019. Filcams CGIL Bologna, Nidil CGIL Bologna, Filt CGIL Bologna v. Deliveroo Italia S.r.l.* https://dbhq-deliveroo-riders-website.cdn.prismic.io/dbhq-deliveroo-riders-website/3f8786d1-5d8c-4d75-b376-94537c6d1634_15605805s+%281%29.pdf.
- Trist, E. (1981). *The evolution of socio-technical systems. Ontario Quality of Working Life Centre, Ontario Ministry of Labour*.
- Tronnier, F., Löbner, S., Azanbayev, A., and Walter, M. (2024). *'A Systematic Literature Review on Gender Bias in AI – Towards Inclusiveness in Machine Learning' (2024)* *https://aisel.aisnet.org/pacis2024/track01_aibussoc/track01_aibussoc/3*.
- Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *42*(1), 230–265.
- Turing, A. M. (1950). *Machinery and Intelligence*. *59*(236), 433–460.
- United Nations - Human Rights Council. (2011). Guiding Principles on Business and Human Rights (fn. 17), n. 18.
- UN Human Rights Council (2026), Women's and girls' rights and artificial intelligence and related digital technologies - Report of the Working Group on discrimination against women and girls, A/HRC/62/48.

- UN Women, & WHO. (2023). *Technology-facilitated Violence against Women: Towards a common definition*. https://www.unwomen.org/sites/default/files/2023-03/Expert-Group-Meeting-report-Technology-facilitated-violence-against-women-en.pdf.
- UNESCO. (2023). *Ethical Impact Assessment. A Tool of the Recommendation on the Ethics of Artificial Intelligence*'.
- UNESCO. (2024). *The weight of words. Gendered language and women's participation and positioning in the labor market*. UNESCO. https://doi.org/10.54678/SYJO7943.
- UNFPA. (2021). *Technology-Facilitated Gender-Based Violence: Making All Spaces Safe*. https://www.unfpa.org/publications/technology-facilitated-gender-based-violence-making-all-spaces-safe.
- UNFPA & Wilson Center. (2022). *Technology-facilitated Gender-based Violence. Data and Measurement: Methodology Matters*. https://www.unfpa.org/sites/default/files/resource-pdf/SummaryDataMeasurementMethodologyMatters.pdf.
- United Nations Advisory Body on Artificial Intelligence. (2024). *Governing AI for humanity: Final report*.
- United Nations, Advisory Body on Artificial Intelligence, Governing AI for humanity : final report / Advisory Body on Artificial Intelligence, https://digitallibrary.un.org/record/4062495?v=pdf.
- United Nations Convention on Access to Information, Public Participation in Decision-Making and Access to Justice in Environmental Matters, June 25, 1998, 2161 U.N.T.S. 447. https://unece.org/environment-policy/public-participation/aarhus-convention/text [unece.org], [treaties.un.org].
- Van Kreij, L. (2019). Towards a comprehensive framework for understanding EU enforcement regimes. *European Journal of Risk Regulation, 10*(3), 439-457. https://doi.org/10.1017/err.2019.52.
- Van Schie, G., Candidatu, L., & Huyskes, D. (2025). Motherhood in the Datafied Welfare State: Investigating the Gendered and Racialized Enactment of Citizenship in Dutch Algorithmic Governance. In J. Van Dijck, K. Van Es, A. Helmond, & F. Van Der Vlist, *Governing the Digital Society*. Amsterdam University Press. https://doi.org/10.5117/9789048562718_ch12.
- Vandenbosch, L., Fardouly, J., & Tiggemann, M. (2022). Social media and body image: Recent trends and future directions. *Current Opinion in Psychology*, *45*, 101289. https://doi.org/10.1016/j.copsyc.2021.12.002.
- Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 97–112. https://doi.org/10.9785/cri-2021-220402.
- Veldanda, A. K., Grob, F., Thakur, S., Pearce, H., Tan, B., Karri, R., & Garg, S. (2023). *Are Emily and Greg Still More Employable than Lakisha and Jamal? Investigating Algorithmic Hiring Bias in the Era of ChatGPT* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2310.05135.
- Vera-Gray, F. (2020). Rape Porn, Cultural Harm, and the Law. In K. Ross, I. Bachmann, V. Cardo, S. Moorti, & M. Scarcelli (Eds), *The International Encyclopedia of Gender, Media, and Communication* (1st edn, pp. 1–5). Wiley. https://doi.org/10.1002/9781119429128.iegmc032.

- Vox – Osservatorio Italiano sui Diritti (2026) *Mappa dell'intolleranza 9*. https://www.voxdiritti.it/wp-content/uploads/2026/05/MAPPA-DELLINTOLLERANZA-9.pdf.
- Wachter, S., Mittelstadt, B., & Russell, C. (2021a). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, *41*, 105567. https://doi.org/10.1016/j.clsr.2021.105567.
- Wachter, S., Mittelstadt, B., and Russell, C. (2021b)., *Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law*, 123 W. VA. L. REV. 735, 744.
- Wachter, S. (2024). *Limitations and Loopholes in the E.U. AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond*. 26 Yale Journal of Law and Technology.
- Wachter-Boettcher, S. (2018). *Technically wrong: Sexist apps, biased algorithms, and other threats of toxic tech* (First published as a Norton paperback). W.W. Norton & Company, independent publishers since 1923.
- Wagner, K. G., & Carlson, C. R. (2025). Abused and silenced: How free speech and online platforms fail women. In D. Harp, I. Bachmann, & J. Loke (Eds), *Handbook on Gender and Digital Media* (pp. 243–255). Edward Elgar Publishing. https://doi.org/10.4337/9781035313570.00030.
- Wall Tweedie, J., Smith, A. C. T., & Ordway, C. (2026). Online Abuse, Sport and Women Athletes: A Narrative Review. *Communication & Sport*, 21674795261426282. https://doi.org/10.1177/21674795261426282.
- Weekes, C. J., Storey, J. E., & Pina, A. (2025). Cyberstalking Perpetrators and Their Methods: A Systematic Literature Review. *Trauma, Violence, & Abuse*, 15248380251333411. https://doi.org/10.1177/15248380251333411.
- Weekes, C. J., Storey, J. E., & Pina, A. (2026). Comparing experiences of cyberstalking from victim and perpetrator perspectives. *Psychology, Crime & Law*, 1–20. https://doi.org/10.1080/1068316X.2026.2700773.
- Weerts, H., Xenidis, R., Tarissan, F., Olsen, H. P., & Pechenizkiy, M. (2023). Algorithmic Unfairness through the Lens of EU Non-Discrimination Law: Or Why the Law is not a Decision Tree. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 805–816. https://doi.org/10.1145/3593013.3594044.
- Weinmann, H., Messingschlager, T. V., & Appel, M. (2026). Gender bias in text-to-image generative artificial intelligence: Neglect and stereotypical presentations across three popular platforms. *New Media & Society*, 14614448261435197. https://doi.org/10.1177/14614448261435197.
- West, S. M., Whittake, M., & Crawford, K. (2019). *Discriminating Systems: Gender, Race and Power in AI*. AI Now Institute. https://ainowinstitute.org/publications/discriminating-systems-gender-race-and-power-in-ai-2.
- Westergaard, D., Moseley, P., Sørup, F. K. H., Baldi, P., & Brunak, S. (2019). Population-wide analysis of differences in disease progression patterns in men and women. *Nature Communications*, *10*(1), 666. https://doi.org/10.1038/s41467-019-08475-9.
- Williams, K. (2025). 'There Are No Limits!': AI Undressing Apps and the Normalization of Nonconsensual Intimate Deepfakes. *Violence Against Women*, 10778012251397966. https://doi.org/10.1177/10778012251397966.

- Wilson, C., Sheridan, L., & Garratt-Reed, D. (2022). What is Cyberstalking? A Review of Measurements. *Journal of Interpersonal Violence*, *37*(11–12), NP9763–NP9783. https://doi.org/10.1177/0886260520985489.
- Wohlfarth, A.-M., Bareither, C., Gollub, T., & Stein, B. (2026). The curation of beauty ideals through Instagram's recommender system: An interdisciplinary approach. *Big Data & Society*, *13*(3), 20539517261476310. https://doi.org/10.1177/20539517261476310.
- Wood, C., O'Sullivan, A., Ljungberg, C., & Lundqvist, A. (2025). *Censorship revealed*. CensHERship and The Case For Her. https://static1.squarespace.com/static/66eab2d398417b6acf847064/t/685e69b37db55d3271edef59/1751017917906/Censorship+Revealed+White+Paper+FINAL+June+2025.pdf.
- Woodlock, D., & Harris, B. (2024). 'You have to be really careful': Technology and the abuse of women with intellectual and cognitive disabilities. *Disability & Society*, *39*(4), 953–973. https://doi.org/10.1080/09687599.2022.2114886.
- Wuidar, S., Engels, L., Lisein, O., Zanoni, P., Rondeaux, G., & Pirsoul, T. (2026). *Recruitment and selection in the age of AI: AI use, awareness of bias risks and recommendations regarding the risks of gender discrimination*. Belgian Institute for the Equality of Women and Men. https://igvm-iefh.belgium.be/sites/default/files/media/documents/AI%20%26%20recruitment%20policy%20brief_EN.pdf.
- Xenidis, R. (2024). *Beyond bias: Algorithmic machines, discrimination law and the analogy trap Transnational Legal Theory*, 1-35. https://doi.org/10.1080/20414005.2024.2307200.
- Xenidis, R. (2017). Shaking the normative foundations of EU equality law: Evolution and hierarchy between market integration and human rights rationales (EUI Department of Law Research Paper No. 2017/04). European University Institute. https://doi.org/10.2139/ssrn.2925110.
- Yeung, K. (2019). *Responsibility and AI - A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework (Study DGI)2019)05).* Council of Europe. https://rm.coe.int/responsability-and-ai-en/168097d9c5.
- Pi, Y., & Proctor, M.1 (2025). Toward empowering ai governance with redress mechanisms. In Cambridge Forum on AI: Law and Governance, volume 1, page e24. Cambridge University Press, 2025.
- Zanfir-Fortuna, G. (2026). Definitions in the AI Act – Main Actors and Key Concepts (Article 3). In G. Malgieri, G. G. Fuster, A. Mantelero, & G. Zanfir-Fortuna (Eds), *The EU Artificial Intelligence Act: A Thematic Commentary* (1st edn, pp. 33–48). Hart Publishing. https://doi.org/10.5040/9781509988570.
- Zeinert, P., Inie, N., & Derczynski, L. (2021). Annotating Online Misogyny. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3181–3197. https://doi.org/10.18653/v1/2021.acl-long.247.
- Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning Fair Representations. Learning fair representations | Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28.
- Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). *Mitigating Unwanted Biases with Adversarial Learning*. https://doi.org/10.48550/arXiv.1801.07593.

- Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 15–20. https://doi.org/10.18653/v1/N18-2003.
- Žliobaitė, I., & Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, *24*(2), 183–201. https://doi.org/10.1007/s10506-016-9182-5.
- Zucker, I., & Prendergast, B. J. (2020). Sex differences in pharmacokinetics predict adverse drug reactions in women. *Biology of Sex Differences*, *11*(1), 32. https://doi.org/10.1186/s13293-020-00308-5.

# ANNEX 1: METHODOLOGICAL APPROACH

The methodological approach of this study was chosen to analyse, at the EU level, the nature, drivers, manifestations and consequences of gender bias in AI sufficiently thorough and efficient. The study aims to provide a robust, interdisciplinary understanding of gender bias in the EU, with due attention to the specific manifestation of TFGBV, and to inform the development of evidence-based recommendations for EU policymakers.

The study consists of three core methodological components:

- desk research through a literature review across law, computer science, social sciences and gender studies;
- targeted EU and national level stakeholder consultation, comprising of a questionnaire; and
- expert review of the draft chapters.

All three methods are qualitative. No quantitative methods have been used for this study because such methods would not adequately lead to the information necessary for the analysis. The chosen qualitative, multi-method research design enabled the integration of diverse sources of evidence. It ensured that the study's findings are based on the EU's legal and policy landscape and on stakeholders' experiences and expertise, and that its recommendations are grounded in real-world practices across the EU.

## a. Desk research

The first phase of research involved desk research addressing the following topics:

- The nature, drivers and manifestations of gender bias in AI;
- The impact of AI on women across key domains of life, including TFGBV;
- The EU legal, policy and governance frameworks and their capacity to prevent, mitigate and redress algorithmic harms.

The review covered EU and international legislation and case law, academic literature, legal and policy documents from EU institutions and agencies, international organisations and civil society and investigative journalism.

Sources were selected for their relevance to the EU, with priority given to evidence from EU Member States, whilst being mindful to reflect variations in size and regional context including South-North and East-West divisions. The study includes research findings, real-world examples and best practices from twenty-two Member States including Austria, Belgium, Bulgaria, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, the Netherlands, Poland, Portugal, Romania, Slovakia, Spain, Sweden. Evidence from other jurisdictions, including the United States and the United Kingdom, was included where comparable EU evidence was lacking.

The literature selection sought to balance authoritative foundational sources with recent studies to ensure both conceptual rigour and topical relevance, while acknowledging the methodological challenges associated with conducting research in a rapidly evolving field.

## b. Stakeholder consultations

To supplement and validate the desk-based review, a targeted stakeholder consultation was conducted, comprising of structured questionnaires.

The comprehensive questionnaire (Annex 2) was distributed in the beginning of September 2026 to institutional actors, experts and civil society organisations. It was also disseminated through the Equinet Secretariat to European equality bodies and gender equality commissions, which were additionally offered the opportunity to complete a simplified version of the questionnaire (Annex 3). Written responses were received from 15 stakeholders including 9 national equality bodies and gender equality commissions from Belgium, Croatia, Estonia, Ireland, Lithuania, Malta, the Netherlands, Spain and Serbia, as well as six EU-level and international organisations and experts, including the EIGE, a representative of the EP, the Council of Europe, CDT Europe, Equinet, and ENNHRI. Annex 2 and 3 provide the consultation questionnaires.

The consultation process served three primary objectives: (i) to identify practices, experiences, and cases that had not been documented in the published literature; (ii) to test and refine emerging findings; and (iii) to inform the development of the recommendations presented in Chapter 8.

### c. Limitations

Several limitations should be acknowledged. First, the stakeholder consultation is not representative. Participants were selected or self-selected, equality bodies are comparatively well represented due to dissemination through the Equinet Secretariat, and responses do not cover all EU Member States. Second, the study did not directly consult individuals affected by gender bias in AI and their experiences are therefore reflected through secondary sources. Third, the evidence base on gender bias in AI remains geographically concentrated and often lacks disaggregated data beyond binary gender categories. Furthermore, the opacity and proprietary nature of many AI systems frequently necessitate reliance on investigative reporting, external audits, or reconstructed models rather than direct access to the systems themselves. Moreover, budget constraints prevented remuneration of participating CSOs and academics, which may have affected participation rates and reflects broader resource asymmetries in policy consultation ecosystems. Finally, given the rapid pace of regulatory, technological, and societal developments in this field, developments occurring after September 2026 fall outside the scope of this study.

# ANNEX 2: CONSULTATION QUESTIONNAIRE ON GENDER BIAS AND ARTIFICIAL INTELLIGENCE

**Section 1 – Legislation, redress and accountability**

**1.** To what extent do the AI Act, the DSA, the GDPR, the Representative Actions Directive, and other relevant legislation adequately address (intersectional) gender bias, both in public and private sector applications?

a. Where do gaps remain for each instrument individually?

b. Where do gaps arise from the interaction between instruments?

**2.** To what extent do existing legal and regulatory frameworks, including their implementation, enable effective attribution of responsibility for (intersectional) gender-biased outcomes in AI/ADM systems and TFGBV across developers, deployers, platforms, and Member States, and how can identified accountability gaps best be remedied?

**3.** How do information asymmetries, difficulties of evidencing harm and cross-border enforcement respectively affect the effectiveness of each type of redress and remedy mechanism (civil and criminal proceedings, DSA and AI Act complaint mechanisms, out-of-court dispute resolution, and platform-level appeals) and which promising practices at European and/or national level address each of these barriers?

**Section 2 – Socio-technical mitigation measures**

**1.** What promising practices exist in the EU for addressing (intersectional) gender bias across the AI lifecycle, from data collection and model design to testing, deployment, and monitoring?

**2.** Which tools and methods (such as bias audits, fairness metrics, (intersectional) gender impact assessments, or diverse testing panels) have proven effective in detecting (intersectional) gender bias in AI systems?

**3.** How can the fundamental tension between audit methodologies and transparency requirements on the one hand, and providers' intellectual property and trade secrets on the other, best be resolved and which existing or emerging legal safeguards or practices show promise here?

**4.** Which approaches by online platforms (such as safety-by-design features, rapid takedown procedures, hash-matching for intimate images, or trusted flagger schemes) have been effective in addressing online discrimination, hate speech and TFGBV, and which promising practices do you identify?

**Section 3 – AI literacy**

**1.** Does AI literacy differ by gender and intersecting characteristics such as ethnicity, socio-economic status, age, and disability? If so, what explains these differences, and what are their effects on risk exposure, contestability of algorithmic outcomes, and participation in AI? If evidence is mixed or inconclusive, please indicate this.

**2.** What initiatives have successfully increased women's participation in the AI ecosystem, and what made them effective?

**Section 4 – Role of civil society**

**1.** What role do civil society organisations, victim support services, and affected communities currently play in detecting, reporting, and remedying (intersectional) gender bias in AI/ADM systems and TFGBV and how can their capacity and involvement be strengthened?

**2.** Which partnerships between civil society, platforms, regulators, and researchers have proven effective in this context?

**Section 5 – Recommendations and priorities**

**1.** Are there any developments, initiatives, or priorities regarding either EU policymaking or your own work on AI and gender that you expect to be relevant for this study?

**2.** What are the key evidence-based recommendations for the European Parliament's FEMM Committee and other relevant EU policymakers regarding (intersectional) gender bias in AI/ADM systems and TFGBV?

**3.** Which priority actions should the FEMM Committee pursue first, for example legislative amendments, oversight mechanisms, or funding initiatives, and what makes these the most urgent?

# ANNEX 3: SIMPLIFIED VERSION OF THE QUESTIONNAIRE

**Query on EP FEMM Study on 'Artificial Intelligence and Gender Bias'**

The authors of a forthcoming study for the **European Parliament's Committee on Women's Rights and Gender Equality (FEMM) on 'Artificial Intelligence and Gender Bias'** have reached out to Equinet. Equality Bodies are warmly invited to share their expertise and input, which would considerably strengthen the study. This study will inform an own-initiative report of the FEMM Committee.

Given the key role of (Gender) Equality Bodies in promoting gender equality in Europe, the authors would greatly appreciate your input on:

1. Any (ongoing) national/regional/local cases involving gender bias in AI or digital technologies, including cases of technology-facilitated gender-based violence, that your organisation is aware of (whether involving the Equality Body directly or not);
2. Interesting best practices or national approaches (including legislation) in this area;
3. Any data or studies on AI literacy, in particular on whether AI literacy differs between genders.

For your information, the study covers:

- Mapping the main forms and sources of gender bias in AI systems used in the EU;
- Analysing how gender bias affects women and girls, including through an intersectional lens, across sectors;
- Emerging forms of gender-based violence in digital spaces;
- Overviewing the relevant EU legal, regulatory and policy framework, including implementation challenges;
- Assessing technical, organisational and governance measures to prevent and mitigate gender bias;
- Identifying accountability, oversight and redress mechanisms and their effectiveness;
- Drafting recommendations for EU policymakers, particularly the FEMM Committee.

This study analyses gender bias in AI and examines emerging forms of technology-facilitated gender-based violence through an intersectional lens, examining its manifestations and impacts on women across domains. The study identifies legal, governance and technical responses, and assesses their effectiveness in preventing and providing redress for algorithmic gender discrimination and technology-facilitated gender-based violence.

This document was provided by Policy Department Citizens, Equality and Culture at the request of the FEMM Committee.

**Disclaimer and copyright.** The opinions expressed in this publication are those of the author(s) only and should not be considered as representative of the European Parliament's official position. The reuse of this document is authorised under a Creative Commons Attribution 4.0 International (CC-BY 4.0) licence (https://creativecommons.org/licenses/by/4.0/). To use or reproduce elements that are not owned by the European Union, permission may need to be sought directly from the respective rightsholders. © European Union, 2026.
© Image on the cover page: Adobe Stock

Administrator responsible: Georgiana SANDU Editorial assistant: Dana AXANEI
Contact: poldep-iust-a@europarl.europa.eu

This document is available on the internet at: www.europarl.europa.eu/supporting-analyses
IUST-2026-A-FEMM-IC-044
Print ISBN 978-92-848-4198-1 | doi: 10.2861/6259540 | QA-01-26-257-EN-C
PDF ISBN 978-92-848-4197-4 | doi: 10.2861/3567357 | QA-01-26-257-EN-N