

D6.2

Social, legal and policy landscapes of AI-fairness 2nd version

Contact us

www.aequitas-project.eu

info@aequitas-project.eu

Deliverable 6.2

Social, legal and policy landscapes of AI-fairness 2nd version

DELIVERABLE TYPE

Report

MONTH AND DATE OF DELIVERY

Month 26, December 31, 2024

WORK PACKAGE

WP 6

LEADER

ALLAI

DISSEMINATION LEVEL

Public

AUTHORS

Catelijne Muller

Programme

Contract Number

Duration

Start

Horizon
Europe

101070363

36 Months

November
1, 2022

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf.

Contributors

NAME	ORGANISATION
Laura Sartori	Alma Mater Studiorum – Università di Bologna
Catelijne Muller	ALLAI
Imane Hmiddou	ALLAI

Peer Reviews

NAME	ORGANISATION
Paul Lemmens	PRE
Andrea Borghesi	UNIBO

Revision History

VERSION	DATE	REVIEWER	MODIFICATIONS
0.1	28/11/2024	Catelijne Muller	Outline and first draft
0.2	10/12/2024	Catelijne Muller	First version shared with reviewers
0.3	15/12/2024	PL and AB	Revisions and comments provided
0.4	21/12/2024	Catelijne Muller	Document revised
0.5	25/12/2024	Laura Sartori	Integration social landscape

Table of Abbreviations and Acronyms

Abbreviation	Meaning
AI	Artificial Intelligence
AI Act	Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts
AI Convention	Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law
EGTAI	Ethics Guidelines for Trustworthy AI of the EU High Level Expert Group on Artificial Intelligence (2019)
EU	European Union
WP	Work Package

Index of Contents

Index of Contents.....	5
Index of Figures.....	6
1 Executive Summary.....	7
2 Introduction.....	8
3 Social landscape of AI Fairness.....	8
4 Ethical Landscape of AI Fairness.....	10
5 Legal Landscape of AI Fairness.....	12
5.1 EU AI Act.....	12
5.2 Objective.....	12
5.3 AI system definition.....	13
5.4 Scope	13
5.5 Structure – risk-based approach	13
5.6 Prohibited AI practices	14
5.7 High-risk AI systems	15
5.8 Actors	16
5.9 Requirements and obligations that relate to AI fairness	17
6. European AI Convention	19
6.1 Aim and Scope of the AI Convention.....	19
6.2 The AI definition.....	21
6.3 General Obligations and the Pursuit of Fairness	21
6.4 Specific Measures for Equality and Non-Discrimination	21
6.5 Counter Mechanisms	22
7. Legal notions of AI-Fairness in HR, Recruiting and Candidate Selection.....	23
7.1 Regulatory instruments	23
7.2 Self-regulatory instrument.....	23
8. Legal notions of AI-Fairness in Healthcare	24
9. Legal notions of AI-Fairness regarding disadvantaged groups	24
10 Policy Developments around AI-Fairness	25
11 Conclusion.....	25

Index of Figures

Figure 1 Risk pyramid to illustrate the risk-based approach of the AI Act. **Errore. Il segnalibro non è definito.**

1 Executive Summary

This Deliverable aims to build on and update the preliminary social, ethical and legal landscape of AI fairness of Deliverable 6.1 by delving into the main regulatory and policy developments relevant to AI fairness since the presentation of D6.1.

Deliverable 6.1 already extensively elaborated on the ethical landscape by delving into the fairness aspects of the EU Ethics Guidelines for Trustworthy AI. WP6 analysed EGTAI and found that elements of AI-Fairness are engrained throughout it (both as a main ethical principle as well as in each of the 7 key requirements for Trustworthy AI of EGTAI, endorsed by the European Commission).

The social landscape of AI fairness highlights the importance of considering cultural, historical and institutional contexts when defining fairness and addressing inequalities. Challenges such as biased training data and inaccessible technologies often marginalise underrepresented groups, including women and LGBTQ+ individuals. Addressing these issues requires the adoption of social notions of AI fairness, which emphasise the need for participatory approaches that amplify marginalised voices and ensure that AI systems promote equity and inclusivity, rather than perpetuate systemic biases. A socio-technical framework, such as the socio-technical matrix proposed in Deliverable 6.4, enables a comprehensive assessment of AI systems that engages diverse stakeholders to promote fairness and mitigate inequalities.

The EU AI Act and the Council of Europe's AI Convention have been developed and concluded over the course of this project and is currently in its implementation phase. This Deliverable will provide an analysis of the relevant elements related to AI fairness of these regulatory instruments.

The ethical, legal and social elements identified through this Deliverable have been continuously integrated into various parts of the project. Particularly several sub-methodologies of the Fair-by-Design Methodology of Deliverable 6.4. This Deliverable also clearly references which elements were implemented where.

2 Introduction

This deliverable serves as a continuation and expansion of the foundational work established in Deliverable 6.1, which examined the preliminary social, ethical, and legal dimensions of AI fairness. Recognizing the rapid evolution of artificial intelligence and its profound societal impact, this report focuses on the significant regulatory and policy advancements that have emerged since the completion of D6.1. By addressing these developments, the document aims to provide a current and thorough analysis of AI fairness within the broader ethical framework established by the European Union.

Building on the ethical principles articulated in the EU Ethics Guidelines for Trustworthy AI (EGTAI), this deliverable explores fairness as both a fundamental ethical tenet and a critical component of the seven key requirements for Trustworthy AI. It delves into the intersection of fairness with emerging regulatory frameworks, assessing their implications for developers, policymakers, and society at large. In doing so, it seeks to bridge theoretical considerations with practical applications, ensuring that fairness remains at the forefront of AI development and governance.

The structure of this deliverable is designed to guide the reader through a comprehensive exploration of the subject. Beginning with a summary of the advancements in AI fairness since D6.1, the document transitions into an in-depth discussion of the evolving policy landscape, highlighting key regulatory initiatives and their alignment with fairness principles. The analysis further addresses the challenges and opportunities presented by these developments, offering actionable insights to promote fairness in AI systems.

Ultimately, this deliverable aspires to contribute to the ongoing dialogue surrounding the ethical, legal, and social dimensions of AI, reinforcing the commitment to a fair, inclusive, and trustworthy digital future.

3 Social landscape of AI Fairness

Since AI systems have become pervasive across public and private sectors, discussions about algorithmic fairness have gained prominence. The social landscape of AI fairness refers to the broader societal context in which fairness in AI systems is considered. It includes how AI systems impact different social groups, how biases in training data or algorithms reflect systemic inequalities and the ethical, regulatory, and public responses to these issues. Deliverable 5.1 has extensively explored the concept of fairness through a sociotechnical lens that sustains the acknowledgment of mutual influences between technical and social structures in line with the approach of Science and Technology Studies (STS). STS adopts an analytical perspective that views scientific and technological outcomes—conceptualised as "technoscience"—as the result of social processes, best understood through approaches and methodologies that highlight their practical and context-specific nature. The advantage of such an approach is that it

emphasizes the culture and methods of knowledge creation in computer science. Following the critical work of notable predecessors from the Nineties, like Lucy Suchman and Phil Agre, our approach moves from a solution-oriented approach to a process-oriented one and advocates for including social actors, institutions, and interactions alongside technology.

A sociotechnical assessment of AI-based solutions helps identify and evaluate the assumptions, epistemological foundations, and implicit claims underlying these technologies. A key point is that if a system relies on biased, unfair, or pseudoscientific foundations, its functionality alone cannot justify its deployment. Therefore, Deliverable 6.4 offers a sociotechnical matrix assessment framework. Inspired by Sloane et al. (2022) work¹, this framework provides a systematic approach to ensure AI solutions rely on solid theoretical and practical grounds. A simplified version of this matrix has been developed to assist the stakeholders involved at this stage in documenting and reflecting on key aspects of their AI solution, facilitating ethical assessment and technical refinement in subsequent stages.

To better understand the social landscape of AI fairness, it is crucial to consider the social notions of AI fairness, which lay the foundation for defining what fairness means in practice. These notions mediate societal disparities, highlighting the role of AI systems in reflecting or challenging such inequalities. Importantly, social notions of fairness are dynamic and context-dependent, shaped by cultural, historical, and institutional factors. They emphasize the need for participatory approaches (See D6.3) that include marginalized voices to ensure AI systems promote equity and inclusivity rather than reinforce systemic biases.

Bias and unfairness may arise at different stages of the AI lifecycle (see Deliverables 5.1 and 6.4). The concept of the “knotted pipeline” of information illustrates the complexity of developing and deploying AI systems, contrasting with the idealized notion of a smooth information flow reflected in a technical system. Challenges such as data quality, curation, standardization, and decision-making introduce barriers that often amplify inequalities. In the workplace, recruitment AI systems are usually trained on historical data, which may reflect existing biases. For instance, if past hiring data favored white male candidates for managerial roles, the algorithm could penalize women or individuals from minority groups, perpetuating gender gaps and, more in general, inequities. Underrepresented groups in training data, such as LGBTQ+ individuals or people with disabilities, may receive lower evaluations compared to other candidates. Additionally, certain groups, such as migrants or individuals with low digital literacy, may struggle to access AI-based recruitment platforms or meet technological requirements (e.g.,

¹ Sloane, M., Moss, E. and Chowdhury, R. (2022). A Silicon Valley love triangle: Hiring algorithms, pseudo-science, and the quest for auditability, in *Patterns*, 3,2, doi: 10.1016/j.patter.2021.100425

automated video interviews) to submit to specific job postings, creating barriers to equal opportunities in job searches.

In the AEQUITAS project, we specifically focus on the social notions of fairness related to marginalized minorities, with particular attention to gender and the LGBTQ+ community. Gender biases refer to patterns of behavior and stereotypes that put women, men, and gender-diverse people in unequal and unjust situations. Social rituals, norms, and expectations reinforce the “gendering” of economic, social, and political practices, often perpetuated by AI systems. A key challenge arises because most fairness approaches rely on observable characteristics (e.g., race, gender), assuming these are present in datasets. However, sexual orientation and gender identity are often unobserved or sensitive, complicating fairness efforts. AI systems might be harmful to LGBTQ+ people in many ways. For example, Automated Gender Recognition (AGR) tools put at risk gender non-conforming individuals, who are often misgendered or flagged as “anomalies.” When these systems are applied in recruitment or border controls, individuals may be compelled to explain their personal history, exposing them to unnecessary risks. Facial recognition cameras equipped with AGR affect fundamental rights, including the freedom of movement, access to public services, and the ability to participate in public protests without fear of threats or attacks.

Viewing AI as a sociotechnical tool rather than a purely technical one reframes algorithmic fairness as a sociotechnical challenge. The “knots” symbolize interconnected obstacles, each contributing to inequality in distinct ways, making them impossible to address in isolation (See D5.1).

A sociotechnical perspective enables a nuanced understanding of the interplay between technology and society (as shown in D2.1), mitigating discrimination and fostering more inclusive systems that challenge inequalities rather than reinforce them.

Therefore, as already outlined in Deliverable 6.1 (Sections 4.1 and 4.2), we define the social landscape of AI fairness as the combination of two processes:

Recording and analyzing social notions of AI fairness and unfairness and its manifestations in society within the various use case domains.

Understanding the level of awareness and knowledge of AI fairness among stakeholders from various domains and creating methods to involve them in the decision-making, design, and development of AI systems (see Integration to Deliverable 2.2).

4 Ethical Landscape of AI Fairness

Deliverable 6.1 extensively elaborated on the ethical landscape by delving into the fairness aspects of the EU Ethics Guidelines for Trustworthy AI. WP6 analysed EGTAI and found that elements of AI-Fairness are engrained throughout it (both as a main

ethical principle as well as in each of the 7 key requirements for Trustworthy AI of EGTAI, endorsed by the European Commission.

We note that EGTAI encapsulates both social notions of AI fairness as well as legal notions of AI fairness, particularly when looking at the recently entered into force EU AI Act. It builds on the 7 requirements for trustworthy AI of the Ethics Guidelines in Recital (27), which explicitly recalls the EGTAI and indicates that the application of the 7 requirements should be translated, when possible, in the design and use of AI models. It further states that they should in any case serve as a basis for the drafting of codes of conduct under this Regulation. All stakeholders, including industry, academia, civil society and standardization organizations, are encouraged to take into account as appropriate the ethical principles for the development of voluntary best practices and standards. Several EGTAI key requirements (or parts thereof) have been inserted in Chapter 2 - Requirements for high-risk AI systems of the AI Act. In light of its aim of the act, to protect health, safety and fundamental rights (the latter including fairness-related fundamental rights) against adverse effects of AI, many of these requirements for AI systems aim to eliminate or mitigate the risk of bias, discrimination, inequality or unfairness. Hence, the AI Act to a certain extent implements and codifies EGTAI and translates the broader ethical notion of 'Diversity, Non-discrimination and Fairness' (EGTAI key requirement #5) into actionable obligations for the AI system's development and use.

The recently adopted AI Convention of the Council of Europe also codifies this broader ethical notion of fairness by introducing specific measures for equality and non-discrimination.

While these are important developments, we strongly believe that EGTAI and its key requirements will continue to play a crucial role in achieving AI fairness, for example by serving as a tool for regulatory interpretation or for addressing regulatory uncertainties or gaps.

Apart from that, we introduced EGTAI in one of the sub-methodologies of Deliverable 6.4.: The Trustworthy AI Readiness Assessment. This methodology's primary purpose is to guide stakeholders in systematically evaluating AI systems' readiness while foregrounding ethical, social, and legal responsibilities. This evaluation serves as a diagnostic tool to identify the **ethical**, **sociotechnical**, and **legal** principles with which the system must align before proceeding with development.

The framework consists of seven sequential steps², whereby EGTAI is leveraged in step 4 "Ethical Assessment". Throughout these steps, particular attention is paid to

² The 7 steps are: 1. **Problem Definition**, where the scope and context of the AI application are established; 2. **Solution Definition**, involving the development of potential AI and non-AI interventions; 3. **Stakeholder Identification**, mapping relevant stakeholders and factors influencing the problem; 4. **Ethical Assessment**, evaluating solutions against the EU Ethics Guidelines for Trustworthy AI; 5. **Socio-Ethical Lens**, applying socio-ethical assessments; 6.

addressing potential bias traps and effects identified in social sciences literature, including **portability trap**, **framing trap**, **formalism trap**, and **ripple effect** considerations, ensuring a comprehensive evaluation of both ethical and social dimensions.

5 Legal Landscape of AI Fairness

During AEQUITAS two prominent regulatory instruments were concluded that changed the AI regulatory landscape significantly: the EU AI Act and the European Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. Both instruments have serious implications for the development and use of AI in Europe and for achieving fairness in AI.

5.1 EU AI Act

The legislative process for the AI Act was officially finalized this summer, resulting in the AI Act having entered into force in August 2024. Its implementation process is however gradual and spans a period of almost 7 years.

5.2 Objective

The AI Act is based on EU values and fundamental rights and aims to give people the confidence to embrace AI-based solutions, while encouraging businesses to develop them. According to the Commission, the use of AI with its specific characteristics (e.g., opacity, complexity, dependency on data, autonomous behavior) can adversely affect several fundamental rights enshrined in the European Charter of Fundamental Rights (ECFR). Hence, the objective of the AI Act is to protect health, safety and fundamental rights from the ill effects of AI.

Fairness related rights in the EU Charter of Fundamental Rights

The ECFR holds an entire Title related to “Equality” that includes various fairness-related rights. Accordingly, these fundamental rights of the ECFR include:

- Freedom of thought, conscience and religion, art. 10
- Equality before the law, art. 20
- Non-discrimination, art. 21
- Cultural, religious and linguistic diversity, art. 22
- Equality between women and men, art. 23
- The rights of the child, art. 24
- The rights of the elderly, art. 25

Solution Selection, choosing the most viable solution based on comprehensive evaluation; and 7. **Reflection and Documentation**, reviewing the assessment process and documenting findings in the AEQUITAS Trustworthy AI Leaflet.

- Integration of persons with disabilities, art. 26

Concomitantly, other fundamental rights could be affected because of infringement of the above equality, non-discrimination and diversity rights, such as:

- Right to life, liberty and security, art. 2 and art. 6
- Fair and just working conditions, art. 31
- Presumption of innocence, art. 48
- Freedom to choose an occupation, art. 35
- Access rights to:
 - Education, art. 14
 - Social Security, art. 34
 - Healthcare, art. 35
 - Services of general economic interest, art. 36

5.3 AI system definition

EU lawmakers agreed on (largely) adopting the OECD definition of AI³. 'AI system' means:

"A machine-based system designed to operate with varying levels of autonomy, 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."

5.4 Scope

Article 2 specifies conditions under which certain domains and stakeholders, such as the military, defence, national security, and scientific research and development, are not required to comply with the AIA. any research, testing or development activity regarding AI systems or models prior to their being placed on the market or put into service. It should however be noted that many of the technical requirements of the AI Act need to be 'built in' at the early (i.e. research) stages of development and many of the procedural requirements can only be met if already considered during the research phase.

5.5 Structure – risk-based approach

The AI Act is a horizontal hybrid regulation (product/fundamental rights protection) and takes a risk-based approach: the higher the risk, the stricter the rule. This approach can best be illustrated as a 'risk-pyramid'.

³ [OECD definition](#)

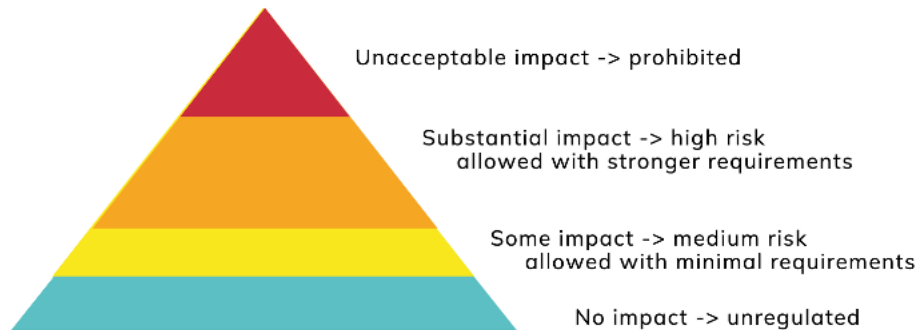


Figura 1- Risk pyramid to illustrate the risk-based approach of the AI Act

The top of the pyramid depicts AI practices that pose such a large risk to health, safety or fundamental rights, that they will not be allowed in the EU.

One step down from the top of the risk pyramid are the AI-systems that are considered high-risk. According to the Commission, these systems do pose a high risk to health, safety, and fundamental rights, but bring enough social benefit to justify their use, provided that multiple requirements are met.

For AI systems with limited impact on health, safety and fundamental rights only transparency requirements apply (art. 50 AI Act).

The AI Act holds a separate regime for so-called “General Purpose AI Models” (GPAI models) such as LLM’s, including certain transparency measures and copyright protection requirements, and more stringent requirements for GPAI models with a systemic risk. It should also be noted, that when using a GPAI model as part of an AI system with a determined purpose, the risk-based regime will apply.

5.6 Prohibited AI practices

Of the prohibited AI practices described in Art. 5 of the AI Act the practice of social scoring is particularly relevant for AEQUITAS.

Art. 5.1(c) prohibits the placing on the market, the putting into service or the use of AI systems for the evaluation or classification of natural persons or groups of persons over a certain period of time based on their social behaviour or known, inferred or predicted personal or personality characteristics, with the social score leading to either or both of the following:

- (i) detrimental or unfavourable treatment of certain natural persons or groups of persons in social contexts that are unrelated to the contexts in which the data was originally generated or collected;

(ii) detrimental or unfavourable treatment of certain natural persons or groups of persons that is unjustified or disproportionate to their social behaviour or its gravity;

Recital 31 clarifies that *‘AI systems providing social scoring of natural persons by public or private actors may lead to discriminatory outcomes and the exclusion of certain groups.*

Social scoring is often perceived as a rare activity that perhaps is employed in China. However, many AI-systems that assess or evaluate people (e.g. several domains listed in ANNEX III relevant for AEQUITAS use data on (or data that infers or predicts)) social behaviour or personal/personality characteristics that are from another context, or unjustified or disproportionate given the assessed/evaluated behaviour or its gravity. If this use leads to unfavourable or detrimental treatment such as discriminatory outcomes or exclusion of certain groups, the use of the AI-system could be considered social scoring.

For example, financial behaviour (credit scoring) would be considered unjustified in sectors such as employment, education, access to healthcare and social services.

In the same spirit, behaviour or personal/personality characteristics inferred from online behaviour (on social media, search engines, website visits, etc.) would be considered unjustified, irrelevant, arbitrary, subjective, or otherwise unreasonable in contexts such as employment, education, access to healthcare or social services.

Using such information to remove someone’s resume from the stack of applicants, or deny someone entry into a study or school, would be considered social scoring.

But also flagging someone to be of low creditworthiness (and denying a loan) based on quotidian actions, such as missing a single payment, would be considered disproportionate and considered social scoring.

Deliverable 6.4 includes a sub-methodology to aid assessing whether an AI practice could be considered social scoring.

5.7 High-risk AI systems

The AI Act considers AI that is a safety component of a harmonized product listed in ANNEX II part A, or is itself such a product, and that already must undergo a third-party assessment pursuant to harmonized EU regulation, to be high risk (art. 6).

The AI Act in ANNEX III also lists several AI-systems in a total of 8 areas that are considered high-risk. These are called ‘standalone high-risk AI’ and relevant areas for AEQUITAS’ uses cases on this list are biometric identification and categorization, employment (AI systems intended to be used for the recruitment or selection of natural persons (...)) and access and enjoyment of (...) essential public services, including education and healthcare.

In exceptional cases, AI systems that are listed in ANNEX III, but have no significant impact on health, safety or fundamental rights, can be exempted, but only under limited circumstances (art. 6.3 AI Act). AI systems that fall within the above categories need to satisfy a large number of requirements before they can be put on the European internal market. These requirements involve risk management, data quality and governance, technical documentation and record-keeping, transparency and interpretability, human oversight, accuracy, robustness and cybersecurity, all to be met through prior conformity assessment. Moreover, numerous procedural requirements apply, depending on the actor developing, deploying or using the AI system.

5.8 Actors

The AI Act includes a complex system of obligations for different actors along the AI supply chain (Chapter II Section 3), from which the following actors most relevant for healthcare can be identified:

- AI-provider: natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge;
- AI-deployer: ‘deployer’ means a natural or legal person, public authority, agency or other body using an AI system under its authority (except where the AI system is used in the course of personal non-professional activity).

Other roles defined in the AI Act are ‘authorised representative’, ‘importer’, ‘distributor’, ‘product manufacturer’ and ‘operator’ the latter referring to the collective of these actors.

For each of these actors, various obligations apply, so proper identification of such roles is crucial to understand the implications of the AI Act. This is even more important when an actor (irrespective of their role) that places a high-risk AI system on the market or puts it into service under its name or trademark, modifies the intended purpose of a high-risk AI system or makes a substantial modification to an AI system that turns it into a high-risk AI system, effectively becomes an AI-provider and thus inherits the obligations of the AI-provider (art. 25).

An AI provider (or any actor that effectively becomes a provider pursuant to art. 25) generally has the most obligations under the AI Act, predominantly:

- Ensure compliance with the requirements for high-risk AI of Title III, Chapter 2 of the AI Act:
 - Risk management system (art. 9)
 - Data and data governance (art. 10)
 - Technical documentation (art. 11)
 - Record-keeping (art. 12)
 - Transparency and provision of information to users (art. 13)

- Human oversight (art. 14)
- Accuracy, robustness, and cybersecurity (art. 15)
- Set up a quality management system that includes (art. 17)
- Draw up and keep available technical documentation (art. 18)
- Keep logs automatically generated by the high-risk AI system (art. 19)
- (Have) perform the prior conformity assessment and draw up the EU declaration of conformity (art. 43, art. 47)
- EU database registration (art. 49(1))
- Take corrective actions and duty of information (art. 20)
- Collaborate with competent authorities (art. 21)
- Affix CE marking (art. 48)

AI deployers that have not put a high-risk AI system on the market or into service, but use it under their professional authority (for example having purchased or procured an AI system or license) have specific obligations throughout the AI lifecycle (art. 26):

- Follow the instructions of use set by the provider
- Assign human oversight to competent, trained, supported and authorized natural persons (art. 14)
- Monitor the AI system in operation (art. 72)
- Ensure relevant and sufficiently representative input data
- Monitor and report serious incidents/malfunctioning
- Keep automated logs
- Inform workers' representatives and affected workers
- Cooperate with national competent authorities
- Perform a Fundamental Rights Impact Assessment (FRIA) (art. 27)

5.9 Requirements and obligations that relate to AI fairness

As explained above, many requirements and obligations under the AI Act are aimed at effectuating the protection of health, safety and fundamental rights, including the earlier mentioned fairness-related fundamental rights. WP6 has identified and analyzed these requirements and obligations and translated them into practical actions at the data, model and outcome interpretation level, to aid the development of the 3 AEQUITAS engines. Below is an overview and analysis of the relevant requirements and obligations for high-risk AI systems.

Risk management system (art. 9)

Art. 9 requires the establishment of a risk management system that includes “the identification and analysis of the known and the reasonably foreseeable risks that the high-risk AI system can pose to health, safety or **fundamental rights** when the high-risk AI system is used in accordance with its intended purpose;”.

Deliverable 6.4 includes a sub-methodology to aid in conducting a Fundamental Rights Impact Assessment for Fairness, focusing on the AEQUITAS HR use case of recruiting.

Data and data governance (art. 10)

Art 10 lays down stringent requirements for data quality and data governance that are largely related to eliminating or mitigating biased outcomes due to biased training, testing and validation data(sets).

Art 10(3) demands these data(sets) to be relevant, sufficiently representative, and to the best extent possible, free of errors and complete in view of the intended purpose. WP6, in close collaboration with other WPs, found that all these elements are important to assess both at dataset as well as at data feature level, to work towards fair training, testing and validation data.

Article 10(2) demands proper data governance practices that include:

- examination in view of possible biases that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited under Union law, especially where data outputs influence inputs for future operations;
- appropriate measures to detect, prevent and mitigate possible biases identified.

Deliverable 6.4 includes a sub-methodology to aid in assessing and documenting fairness of training, validation and testing data: the Fair Data Collection, Governance and Management methodology.

Transparency (art. 13)

Art. 13(1) requires that high-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers (e.g. healthcare practitioners) to interpret the system's output and use it appropriately, evaluate its functionality, and comprehend its strengths and limitations (through granular instructions of use). Art. 13 hence requires explainable and interpretable AI.

Human oversight (art. 14)

As unfairness in AI can also enter the AI lifecycle through a system's model design or selection, or during the interpretation of a system's outcomes, art. 14 is also relevant. Art. 14 demands that high-risk AI systems be designed and developed in such a way (...) that they can be effectively overseen by natural persons during the entire AI lifecycle. It also prescribes that human oversight shall aim to prevent or minimize the risks to health, safety or fundamental rights that may emerge.

For effective human oversight, it further prescribes that natural persons should be enabled to do the following:

- Understand the high-risk AI system's capacities and limitations and monitor its operation to detect and address issues.
- Be aware of the tendency to over-rely on the AI system's output, especially for systems providing decision-making information.
- Correctly interpret the AI system's output using available tools and methods.
- Decide not to use the AI system or override its output.
- Intervene in the AI system's operation or stop the system safely.

Both Art. 13 and 14 require high-risk AI systems to be designed in a sufficiently transparent manner, making their output understandable to humans.

Deliverable 6.4 includes a sub-methodology to aid in ensuring fairness when designing or selecting a model: the Fair Model Methodology.

Second, art. 14 requires humans to address and counter their own 'biases' when interpreting a system's outcomes.

Deliverable 6.4 includes a sub-methodology to aid in ensuring fairness when interpreting an AI system's outcomes: the Fair Outcome Interpretation Methodology.

6. European AI Convention

The Council of Europe's AI Convention on Artificial Intelligence, Human Rights, Democracy, and the Rule of Law ("AI Convention") recognizes that AI systems can pose significant risks to virtually all human rights, and particularly to fairness, equality, and non-discrimination. The AI Convention aims to address these risks by setting out general obligations and specific measures that the parties must implement throughout the lifecycle of AI systems. This chapter will delve into how the AI Convention seeks to safeguard fairness-related human rights.

6.1 Aim and Scope of the AI Convention

The AI Convention has several key provisions regarding its aim and scope. The AI Convention has been negotiated by 57 states and aims to have a global impact. The convention's overarching aim is to ensure that AI activities align with human rights, democracy, and the rule of law. The convention does not establish new human rights, but it seeks to reinforce existing international human rights laws and relevant domestic legislation in the context of AI. Parties to the convention are given flexibility in how they meet their obligations, with the option to either adopt new measures or maintain existing ones.

The convention considers all activities within the lifecycle of AI systems, rather than just the AI system itself. The lifecycle very much aligns with the AEQUITAS AI lifecycle (see D6.4) and includes:

- Planning and design
- Data collection and processing
- System development (including fine-tuning)
- Testing, verification, validation
- Deployment
- Operation and monitoring
- Retirement

The convention specifies that measures should be adjusted based on the severity and probability of an AI system's adverse impacts on human rights, democracy, and the rule of law. However, the interpretation and application of these measures should consider all relevant circumstances without pre-set limits.

The convention mandates that AI systems developed or used by public actors, or private actors acting on their behalf, must comply with its provisions. This includes private actors who are delegated responsibility by a public entity, are contracted by a public authority, provide public services, or are involved in public procurement.

For private actors not included in the above category, parties to the convention have two options:

- Apply the convention directly to these private parties.
- Use other measures (administrative or voluntary) to ensure protection, if these measures do not undermine human rights, democracy, and the rule of law. Parties must declare which option they will use upon signing, ratifying, or acceding to the convention.

A federal clause allows for reservations to accommodate restrictions faced by federal states when implementing the convention due to their laws or power distributions between national and regional authorities.

AI systems used in research and development are excluded from the scope of the convention, if they are not yet available for use and do not interfere with human rights, democracy, or the rule of law. However, the convention's principles should be considered during the research and innovation phase, and this exemption does not negate the principle of 'safe innovation' or the need to exchange information on risks and effects.

Parties can choose whether to apply the convention to AI systems related to national security, provided these activities comply with international and domestic laws, and respect democratic processes. The focus on national security means that dual-use AI not related to national security, as well as regular law enforcement activities, remain within the scope of the convention.

AI systems developed or used for national defence are excluded from the scope of the convention. This exclusion is based on the principle that matters relating to national defence do not fall within the purview of the Council of Europe.

6.2 The AI definition

The AI Convention adopts the same definition of AI as the one developed by the OECD, which is further clarified in the "Explanatory memorandum on the updated OECD definition of an AI system". This decision was made to foster international cooperation and consistency in AI governance. While the OECD definition is intended to be legally precise and certain, it is also abstract enough to include systems that could potentially harm human rights, democracy, and the rule of law. The definition should always be read in the context of the other provisions of the AI Convention. This means that even if an AI system falls within the definition, its potential to adversely affect human rights, democracy, and the rule of law determines whether the AI Convention applies to it. The convention is concerned with AI systems that pose a risk to human rights, democracy, and the rule of law rather than all AI systems. Therefore, it is the potential for negative impacts that determines the reach of the convention, rather than just whether a system fits the definition of AI.

6.3 General Obligations and the Pursuit of Fairness

At its core, the AI Convention aims to ensure that activities within the lifecycle of AI systems are fully consistent with human rights, democracy, and the rule of law. This includes a commitment to equality and non-discrimination. Article 4 of the AI Convention obliges the parties to adopt or maintain measures ensuring that AI systems are consistent with human rights obligations, as enshrined in applicable international law and domestic law. This includes all global, Council of Europe, and EU instruments that a state may be a party to.

6.4 Specific Measures for Equality and Non-Discrimination

One of the most reported impacts of AI on human rights is its effect on the prohibition of discrimination and the right to equal treatment. In many instances, AI has been shown to perpetuate, amplify, and possibly enshrine discriminatory or otherwise unacceptable biases. AI can also enlarge the group of impacted people by grouping them based on shared characteristics. Moreover, AI systems can obscure the existence of biases, thereby marginalizing the social control mechanisms that govern human behaviour. The right to non-discrimination is established in a vast body of international and domestic human rights law, which provides a solid legal foundation. Article 10 of the AI Convention specifically refers to this legal framework.

The AI Convention also requires parties to consider the broader concept of bias, which can occur at many stages in the lifecycle of an AI system. This includes biases of the algorithm's developers; biases built into the model; biases inherent in the training datasets or the aggregation or evaluation of data; biases introduced when AI systems are implemented in real-world settings or as they evolve; human automation or confirmation biases; and social biases, where historical or current inequalities in society are not properly accounted for within the lifecycle of an AI system. The measures to address discrimination, bias, and inequality should not be limited to ensuring an individual is not treated less favourably, but they should also lead to the adoption of

measures to overcome historical and structural barriers to gender equality and fair and just treatment of people of certain groups, that tend to be replicated and amplified by AI systems.

6.5 Counter Mechanisms

The AI Convention introduces transparency, oversight, accountability and the option to impose moratoria or bans as essential safeguards due to the complex, opaque, adaptable, and often autonomous nature of many AI systems, which makes them vulnerable to human rights impacts.

Transparency refers to openness and clarity about AI governance, including the AI system itself. This includes details about the system's purpose, limitations, assumptions, design choices, features, underlying models or algorithms, data, training methods, and quality assurance processes. Transparency also encompasses information about risk mitigation efforts and how the AI system's outputs are derived, requiring an understanding of how the AI system could impact human rights, democracy, and the rule of law. Transparency also includes informing individuals about the processing of their information, the level of automation used in consequential decisions, and the risks associated with the AI system. It also involves measures to identify content generated by AI. Transparency includes both "explainability" and "interpretability." Explainability refers to providing information about why an AI system provides certain information, predictions, recommendations, or decisions, which is particularly crucial in sensitive areas where understanding the reasoning behind AI decisions is essential. Interpretability focuses on understanding how an AI system reaches its predictions or decisions. The internal workings, logic, and decision-making processes of AI systems must be understandable and accessible to human users.

Oversight involves a wide range of mechanisms aimed at monitoring, evaluating, and guiding activities within the lifecycle of AI systems. These mechanisms can include legal, policy, and regulatory measures, ethics guidelines, codes of practice, certification programs, oversight bodies, competent authorities, and auditing schemes. Due to the complexity of AI systems, the AI Convention encourages that AI systems be designed and developed to allow for effective and reliable oversight, including human oversight.

The AI Convention includes the principles of accountability and responsibility to ensure that individuals and organizations can be held responsible for the adverse effects of AI systems. These principles are closely linked to transparency, as they cannot exist without it.

The core objective is to ensure that there are mechanisms in place to hold entities accountable when AI systems cause harm. The AI Convention provides various mechanisms that parties can use to implement accountability and responsibility, such as liability and other civil law regimes, administrative procedures, and criminal law regimes. The principles of accountability and responsibility require clear lines of responsibility so that outcomes can be traced back to specific individuals or organizations. This is crucial

for correctly attributing responsibility when AI systems cause harm. Those responsible for AI systems must consider the potential impact of their systems on human rights, democracy, and the rule of law in an anticipatory manner.

The AI Convention requires states to take appropriate measures to ensure the identification, assessment, prevention, and mitigation of risks posed by AI systems to human rights, democracy, and the rule of law. While the AI Convention does not impose direct bans or moratoria on specific AI systems, it obligates states to consider bans or moratoria whenever an AI system is deemed incompatible with human rights, the functioning of democracy, or the rule of law.

7. Legal notions of AI-Fairness in HR, Recruiting and Candidate Selection

Apart from the ECFR, the European Convention on Human Rights, the AI Act and the AI Convention, other regulatory and self-regulatory instruments that hold notions of AI-Fairness are relevant for HR, recruiting and candidate selection.

7.1 Regulatory instruments

- Directive 2010/41/EU demanding equal treatment for men and women engaged in an activity in a self-employed capacity.
- Directive 2000/78/EC against discrimination at work on grounds of religion or belief, disability, age or sexual orientation.
- Regulation on Freedom of Movement of Workers (492/2011/EU) and the Facilitating Directive on Free Movement of Workers (2014/54/EU) demanding equal treatment of all EU nationals in recruiting, prohibiting discrimination based on medical, vocational or other grounds such as language, except when required for the job.
- Racial Equality Directive (2000/43/EC) prohibiting direct or indirect discrimination based on racial or ethnic origin in relation to conditions for access to employment, self-employment and occupation including selection criteria and recruitment condition (unless it is (i) a genuine and determining occupational requirement; (ii) based on positive discrimination):
 - **Direct:** less favourable treatment in a comparable situation on grounds of ethnicity or race
 - **Indirect:** apparently neutral criterion/practice would put a person in a disadvantaged position due to race or ethnicity (unless objectively justified)

7.2 Self-regulatory instrument

The EU Social Partner Agreement on Digitalisation (self-regulation) states that employment of AI systems in the workplace/HR/recruiting should (a.o.):

- follow the principles of fairness, free from unfair bias and discrimination
- be transparent and explicable
- have effective oversight

Moreover, employers should undertake a risk assessment for (a.o.) confirmation bias or cognitive fatigue.

8. Legal notions of AI-Fairness in Healthcare

Apart from the ECFR, The European Convention on Human Rights, and the AI Convention, the following is relevant for AI in Healthcare.

[Gender Access Directive \(2004/113/EC\)](#) demanding equal access to goods and services for all sexes, including public services.

The proposal for the AI Act classifies AI systems intended to be used as a (safety components of a) medical device as regulated in the Medical Devices Regulation and the In-Vitro Diagnostics Regulation as high risk. Separately, the proposal for the AI Act classifies biometric identification and categorization as high-risk. These systems need to comply with the requirements for high-risk AI as described above.

9. Legal notions of AI-Fairness regarding disadvantaged groups

For the use case regarding the *detection of child neglect and abuse*, it is important to note that the AI Act classifies biometric identification and categorization as high-risk. Moreover, the possibility of performing the prohibited practice of social scoring (art. 5.1(c) of the AI Act) should be critically assessed.

For the use case on *access to education for disadvantaged students*, it is important to note that the AI Act classifies AI systems intended to be used for the purpose of determining access or assigning natural persons to educational and vocational training institutions as high risk. Moreover, the possibility of performing the prohibited practice of social scoring (art. 5.1(c) of the AI Act) should be critically assessed.

Also here, the [Gender Access Directive \(2004/113/EC\)](#) demanding equal access to goods and services for all sexes, including public services such as healthcare and education, is relevant.

10 Policy Developments around AI-Fairness

Continuously following policy developments around AI is crucial, as these are currently evolving. WP6 is scanning, observing, analysing and contributing to these developments continuously, through various methods, such as:

- Direct exchanges with European and national policy makers on AI policy and regulation.
- Contributing to consultations on the AI Act and providing recommendations on the interpretation of the AI Act to EU and national policy makers.
- Active participation in the development of the GPAI Code of Practice.
- Active participation in and organization of exchanges with national parliaments and governments on AI policy and AI regulation.
- Active participation in the OECD Expert Group on AI Incidents.
- Active participation in negotiations and deliberations on the AI Convention and the HUDERIA between the member states of the Council of Europe.

Relevant policy developments to mention are:

- The AI Office requested input on the definition and the prohibitions of the AI Act to help clarify any legal uncertainties.
- A first draft of the GPAI Code of Practice has been released.
- The Council of Europe's Committee on AI (CAI) has developed a Human Rights, Democracy and Rule of Law Impact Assessment framework to accompany the AI Convention. Additional work will be done in 2025 to deepen the framework.
- An upcoming EU Directive on AI at Work might be relevant for the AEQUITAS use case in HR, recruiting and candidate selection.
- A Digital Fairness Act is expected to be proposed in 2025.

11 Conclusion

Fairness is a multifaceted concept rooted in legal norms, ethical principles, and social values. Its conditional and context-dependent nature makes it nearly impossible to establish quantitative thresholds or benchmarks for fairness that can be directly translated into technical measures. While the AI Act seeks to mitigate the risks of unfair or discriminatory outcomes, the risk cannot simply 'be computed away'.

Consortium



UMEÅ
UNIVERSITET



AKKODIS



PHILIPS

LOBA®

ALLAI.



EUROCADRES



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA



Funded by the European Union. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

www.aequitas-project.eu
info@aequitas-project.eu



Funded by
the European Union