AEQUITAS
unbias AI

D6.1
Social, Ethical and Legal Landscapes of AI-Fairness

This content is based on research conducted for the AEQUITAS report "Preliminary Social, Ethical and Legal Landscapes of AI-Fairness"

www.aequitas-project.eu

01. Landscapes of AI-Fairness

Why? To align the three AEQUITAS Engines with the social, legal, ethical and policy elements and context related to AI fairness.

Crucial to **fair and trustworthy AI**, responsible innovation usually refers to **anticipation, reflexivity, inclusion, and responsiveness** by thinking systematically about socio-technical and legal affordances, use context, and the conflicting interests of various stakeholders.

Anticipating unfairness in ALL stages of the AI lifecycle

1 Scoping phase
2 Design phase
3 Development phase
4 Deployment and use phase
5 Dismantling phase

AEQUITAS' **overarching methodology** focuses on the benefits of ***harm anticipation***. Soliciting and involving the **intended end-users as well as those eventually impacted by the technology** is critical to anticipate potential harms both during the ***envisioning (model design)*** and the **response *(mitigation and interpretation)*** stages. AEQUITAS will follow this approach to design and develop an anticipatory experimentation environment that will enable experimenting with, exploring and adjusting the fairness levels of an AI tool.

02. Social Landscape of AI Fairness

How are we identifying the social landscape of AI fairness?

AI Unfairness Manifestations Database

We have created a database to track and analyze manifestations of unfairness caused by AI in various domains. The database allows for the categorization of information related to each manifestation of unfairness, including:

- **Technical details** about the AI techniques used, training data, inputs, outputs, and interpretations
- **Sources of unfairness**, such as data, algorithms/models, or interpretations
- **The ethical, legal, and social implications of AI fairness** involved
- **Individuals or groups** negatively and positively affected by the AI unfairness
- **The types of harm resulting from AI unfairness** are documented
- **Relevant existing or upcoming policies** related to the manifestation at hand

AI Stakeholder Identification Methodology (AISIM)

We have developed a preliminary methodology for identifying **relevant stakeholders** to involve in the **design process of AI-systems**. By utilizing a combination of desk research, our own expert knowledge, and resources from previous projects, we have created a **questionnaire** that guides the selection of targeted user groups and stakeholders to be involved in the design process of AI-systems. The questions are divided into three categories to identify:

Affectees	Decision makers	Domain Experts and Users
Stakeholders affected by the AI-system	Stakeholders that have power over the development and deployment of the AI-system	Stakeholders that have information that would aid with the development of a fair AI-system
Identification & categorization of negatively and positively affected stakeholders and the level of impact	**Identification & categorization of stakeholders with power over the development, deployment and governance of the AI-system and their level of involvement**	**Identification & categorization of stakeholders that can aid the development of the AI-system and their level of involvement**

03. Ethical landscape of AI-Fairness

A preliminary overview of the **ethical landscape of AI-Fairness** was obtained using the **Ethics Guidelines of Trustworthy AI ("EGTAI")** developed by the **High-Level Expert Group on AI (HLEG AI)** of the **European Commission**.

Consequently, **elements of fairness** can be found in the **7 Key Requirements (KR) for Trustworthy AI of EGTAI:**

7 Key Requirements
Fairness related to:

1 Human agency and oversight

Ensures that individuals are not unfairly manipulated, deceived, herded or conditioned, or subject to any other unfair outcomes.

Requires that systems are regularly reviewed and audited to ensure they are operating as intended.

2 Technical robustness and safety

Helps to prevent errors or unintended consequences that could result in bias, discrimination and other instances of unfair treatment.

Includes ensuring that AI-systems are reliable, accurate, secure, resilient to attacks, and have a fallback plan.

3 Privacy and Data Governance

Helps protect individuals' personal information and prevent discrimination based on sensitive characteristics such as race, gender, or sexual orientation.

Data should be free of socially constructed biases, inaccuracies, errors or mistakes, that only duly qualified personnel with the competence and need to access individual's data should be allowed to do so, and that any data used to train AI-systems is collected and used in a lawful and ethical manner.

4 Transparency

Allows individuals and organizations to understand how AI-systems make decisions, to identify any issues of bias or discrimination, and to hold entities accountable in the event of AI unfairness.

5 Diversity, Non-discrimination and Fairness

Helps ensure that the system is fair, avoid the presence of any unfair bias but also ensure that the system is accessible and can be used by all regardless of their age, gender, abilities, or characteristics.

Ensures diversity in the design of the AI-system through the participation of diverse stakeholders, including those who may be directly or indirectly affected by the system.

6 Social and Environmental Well-being

Ensures that the development and deployment of AI-systems aligns with broader social and environmental goals, such as sustainable development, reducing poverty and inequality, promoting human rights, and ensuring democracy.

7 Accountability

Ensures that individuals and organizations can be held accountable for any issues of bias or discrimination that may arise from the use of AI-systems.

Includes establishing clear lines of responsibility and liability, minimizing and reporting negative impacts and trade-offs that influence fairness.

04. Legal landscape of AI-Fairness

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

- Several **European and EU Treaties** (including the European Convention on Human Rights, the EU Charter on Fundamental Rights and the Treaty on the EU) and numerous **EU Regulations and Directives and self-regulatory instruments** (e.g., social partner agreements) hold **notions of AI-Fairness relevant for HR, recruiting and candidate selection**.

  Non-discrimination, right to work, freedom of occupation, freedom of religion, right to privacy and family life, gender and racial equality, equal employment, data protection, equal treatment and opportunity, harassment prevention, fair and transparent digital HR practices, vigilance against compromising human dignity, human oversight, and compliance with GDPR.

- The **proposed AI Act** classifies **AI systems intended to be used for recruitment or selection of natural persons, notably for advertising vacancies, screening or filtering applications, evaluating candidates in the course of interviews or tests**, as **high risk**. Separately, the proposal for the AI Act classifies **biometric identification and categorization** as **high risk**.

  Moreover, the **proposal for the AI Act prohibits social scoring by public authorities**, which is seen as the **evaluation or classification of the trustworthiness of persons based on unrelated or irrelevant social behaviour or personal characteristics**, leading to **detrimental treatment of that person**. This prohibition could be relevant **when using AI for HR, recruiting or candidate selection**, especially as it **often involves a type of scoring**.

Legal notions of AI-Fairness in Healthcare

- Several **European and EU Treaties** (including the European Convention on Human Rights, the EU Charter on Fundamental Rights and the Treaty on the EU) and numerous **EU Regulations and Directives** hold **notions of AI-Fairness relevant for the Healthcare domain**.

  Non-discrimination; human dignity; right to private and family life; right to life; right to preventive healthcare and medical treatment; data protection; positive discrimination; racial equality; gender equality.

- The **proposed AI Act** classifies **AI systems intended to be used as safety components of a medical device**, as regulated in the Medical Devices Regulation as **high risk**. Separately, the proposal for the AI Act classifies **biometric identification and categorization** as **high risk**.

Legal notions of AI-Fairness regarding disadvantaged groups

- Several **European and EU Treaties** (including the European Convention on Human Rights, the EU Charter on Fundamental Rights and the Treaty on the EU) and numerous **EU Regulations and Directives** hold **notions of AI-Fairness** relevant for the two use cases that deal with disadvantaged groups:

  (i) detection of child neglect and abuse, and
  (ii) access to education for disadvantaged students.

  Non-discrimination; human dignity; rights of the child; right to private and family life; right to life; right to preventive healthcare and medical treatment; data protection; racial equality; gender equality; freedom of religion

  Non-discrimination; human dignity; right to private and family life; protection; racial equality; gender equality; freedom of religion; rights of the elderly; rights of people with disability; right to education; freedom to choose an occupation; right to engage in work; equal treatment of qualifications.

- The **proposed 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**. Separately, the proposal for the AI Act classifies **biometric identification and categorization** as **high risk**.

Once the **AI Act** is adopted, systems must comply with a large set of requirements before they can be put on the EU internal market. The AI Act proposal categorizes the requirements into the overarching categories of:

(i) risk management
(ii) data and data governance
(iii) technical documentation and record keeping
(iv) transparency and provision of information to users
(v) measures to ensure human oversight
(vi) accuracy, robustness, and cybersecurity.

What's next?

Future policy developments

- EU AI Regulatory Framework
- AI Act
- AI Treaty of the Council of Europe

Development of the AEQUITAS AI Fairness-by-design Methodology

Scoping phase | Design phase | Development phase | Deployment and use phase | Dismantling phase

Building Blocks

- Identification and assessment of manifestations of **AI unfairness** related to the AI-system and application domain at hand.
- Identification of the **legal landscape of AI-Fairness** related to the AI-system and application domain at hand.
- Identification of the **ethical landscape of AI-Fairness** using EGTAI related to the AI-system and the application domain at hand.
- Identification of relevant stakeholders using the **Stakeholder Identification Methodology**.
- Identification and assessment of **ethical, legal and social AI-Fairness elements in collaboration with stakeholders**

We will develop an AI Fairness-by-Design Methodology that covers the ENTIRE LIFECYCLE OF AN AI SYSTEM

Want to know more about AEQUITAS?

Visit our website for more information:
www.aequitas-project.eu

Follow us
@aequitasEU

Contact us
info@aequitas-project.eu

Funded by the European Union

Be the first to get exclusive project updates by subscribing to our newsletter!