AEQUITAS
unbias AI

D5.1

Fair-by-design Sociological, Legal Methodologies

This content is based on research conducted for the AEQUITAS "Fair-by-design Sociological, Legal Methodologies Preliminary Compendium".

www.aequitas-project.eu

01. What is Fairness?

Types of discrimination

Disparate Treatment
Refers to the situation where an individual is intentionally treated differently based on their membership of a marginalized class.

Disparate Impact
Happens when members of a marginalized class are negatively affected more than others when using a formally neutral policy or rule. It is unintentional or indirect discrimination.

To tackle discrimination, different definitions of fairness can be used:

- **Distributive Fairness**
  Making sure the outcomes of a process are FAIR.
- **Procedural Fairness**
  Ensuring the FAIRNESS of the decision-making process that leads to the outcome.

COMPUTATIONAL FAIRNESS OF AI SYSTEMS

What's missing?

1. The social goal for which the model is deployed

Trying to achieve the social and legal goals through modeling leads to simplified decisions that, by themselves, might be accurate and successful from a technical perspective, but can be inaccurate, unsuccessful, or even unlawful from a social and/or legal perspective. Modelling assumes that the social and legal goals can be formulated by a mathematical utility function that depends on decisions and outcomes. By contrast, human decisionmakers usually elaborate their decisions on several outcomes (such as the defendant's well-being, alleviating circumstances, or a family situation).

2. The individuals subject to the decision

When it comes to population, predictions usually refer to a subset of a population: e.g., prisoners, loan applicants, or A-level students. The issue here is how those sub-samples are created and what the underlying mechanisms of entry are, which can entail social hierarchical structures and power relations. The fairness technical discussion can overlook the unfair processes by which individuals entered the subsample of the population.

3. The decision space where the decision makers interact with the model's predictions

Another issue regards the decision space, that is to say, the number of available decisions, not always a yes or no decision (detain or release). The availability of alternative options and their acceptability is usually overlooked in discussions on mathematical definitions of fairness.

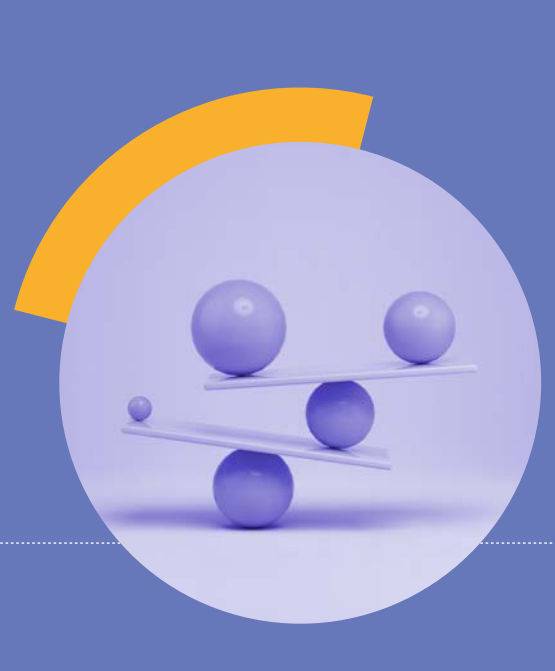
4. The (applicable) legal rules regarding fairness

The EU regulatory framework for AI is being shaped as we speak. Apart from that, AI never operated in a lawless world and many relevant AI-fairness rules relevant to the use cases at hand exist that are usually unknown, and thus overlooked in discussions on mathematical definitions of fairness.

5. The context dependency of the legal rules regarding fairness

The context dependency of legal fairness rules complicates a purely mathematical approach to AI fairness even further, since different circumstances and conditions can result in different interpretations of the rules.

02. Fair-by-Design methodologies



→ Social methodologies

Audits and algorithmic impact assessments

An **audit** can be understood as a comprehensive inspection to check if an algorithmic system is behaving according to rules or norms – this is called a regulatory inspection. A regulatory inspection could be used to assess whether an algorithmic system complied with data protection law, equalities legislation, or insurance industry requirements. This type of inspection would need the participation or cooperation of those deploying the algorithmic system.

Algorithmic Impact Assessments (AIAs) are an emerging method for algorithmic accountability and public trust in AI systems. They provide relevant insights, especially when blended with a wider participatory methodology that encompasses the lived experiences of the people and engaged communities.

Ethnographic approach

Ethnography helps in focusing on local practices and contextual features, typically through in-person interviews, focus groups and observational techniques. On the production of algorithmic systems, there is a rich body of ethnographic work focusing on the technology sector, where ethnographers analyze the role of cultural and organizational processes in shaping the kind of technologies that are built.

Ethnographic approaches shed light on the complex overlap of social, cultural, and technological aspects of computational systems in our daily lives. They provide rich and fine-grained data on how algorithms are built and used. On the production side, ethnographic studies highlight important affinities between workplace cultures and algorithmic design. On the reception side, they show how social practices mediate the uses and actual impact of algorithms.

Focus groups

Focus groups allow a deeper examination of complex issues than other forms of survey research. They are usually used for exploratory research rather than descriptive or explanatory research. They are a useful method for researchers who wish to gather in-depth information about social processes in a specific context.

In combination with surveys, focus groups are useful in getting insights about the perceived unfairness of AI systems as well as details about the sociotechnical imaginaries of AI implications in terms of inequalities.

Cultural and historical critique

Cultural, historical, and economic critique examines the connections between algorithms and the broader structures of social life. Their importance relies on framing the problem from a different alternative perspective, although local practices and contextual features shaping the construction, diffusion, and reception of algorithms are often overlooked.

Participatory methods

Participatory methods include a range of activities with the common goal of enabling ordinary people to play an active and influential part in decisions that affect their lives. In the case of AI, participation is open not only to individuals affected by the act of classification or the final output of an Automated Decision-Making algorithm but also to the many different stakeholders (designers, firms, public administration, associations) of the AI/ML pipeline.
Investing in participatory methods allows for a deeper understanding of the right problem to address, considering what each stakeholder deems important. It also allows for building trust and developing more suitable solutions for the affected individuals and communities. Participatory methods could counterbalance the tendency of computer scientists to focus on the biases in their models and on algorithmic means to solve them.

Survey methods

Survey research is a method involving the use of standardized questionnaires or interviews to collect data about people and their preferences, thoughts, and behaviours in a systematic manner.
This method is best suited for studies that have individual people as the unit of analysis. When applicable to representative samples of the population(s), surveys are really useful to have a precise understanding of the need to overcome discrimination and a qualified representation of the socio-technical imaginaries of the correspondent population.

→ Legal methodologies

↘ **Current Legal Sources**

EU Primary Law	Treaties of the EU, Charter of Fundamental Rights
EU Secondary Law	GDPR, Product Liability Directive, Regulation on the Free Flow of Non-Personal Data, anti-discrimination Directives, Consumer Law, Safety and Health at Work Directives
UN Human Rights Treaties	
Council of Europe Conventions	European Convention on Human Rights
EU Member State Laws	
Domain-specific rules	E.g. Healthcare sector – Medical Device Regulation

↘ **Future Legal Sources**

Recent regulatory developments around AI are laying down technical, social, ethical (and general legal) fairness notions in specific legal requirements.

The EU AI Act	Makes demands that ensure the protection of Fundamental Rights

03. A holistic methodology for Fair-by-Design AI

There is a need for a more holistic approach to fair AI, that includes **technical steps**, **sociological activities**, **legal compliance** measures and understanding, and **ethical considerations**.

The ultimate goal is to design a Fairness-by-Design methodology that integrates technical, legal, ethical and social fairness notions.

The **Fair-by-Design** engine that will be developed within the AEQUITAS project aims to deliver a **practical methodology** that includes all these elements.

BUILDING BLOCKS OF A HOLISTIC METHODOLOGY FOR AI FAIRNESS-BY-DESIGN

- AI-Fairness Impact Assessment/AI-Fairness Readiness Assessment
- Ethics Guidelines for Trustworthy AI and Assessment List for Trustworthy AI (HLEG AI)
- Socio-technical Matrix
- Stakeholder Identification Methodology (developed as part of Deliverable 6.1)
- Stakeholder Engagement Methodology
- Trustworthy AI Deliberation (based on the 7-step exercise for Trustworthy AI, developed for the Trustworthy AI Project (Erasmus+))
- AI Fairness Regulatory Landscape Identification and Assessment
- AI Act Risk Classification
- AI Act High-Risk Requirements Guidance for Fairness
- Fundamental Rights Impact Assessment
- Fair software engineering methodologies architectures and methods:
  - Fair System Architecture Methodology
  - Fairness Criteria Definition, Assessment and Monitoring
  - Fair Data Collection Methodology
- AI-Fairness Evaluation/Bias Audit
- Monitoring in Operation through Critical Control Points

→ What's next?

These **building blocks** will be the base on which we will build the **AEQUITAS Fair-by-Design engine**. Some of them are **socially oriented**, some are **legally oriented**, some are **technically oriented**, and some are a **combination of these orientations**.

In the next stages of the project, we will **further develop these building blocks**, as well as their sub-components and **identify their positions vis-à-vis the AI lifecycle**. We will also identify **their positions vis-à-vis each other**, and their **interplays and overlaps** to determine where and to what extent they could be integrated.

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**!