AEQUITAS

unbias AI

Deliverable 6.1

Preliminary Social, Ethical and Legal Landscapes of AI-Fairness

Contact us

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

Funded by
the European Union

Deliverable 6.1

Preliminary Social, Ethical and Legal Landscapes of AI-Fairness

DELIVERABLE TYPE	MONTH AND DATE OF DELIVERY
Report	Month 03, January 31, 2023

WORK PACKAGE	LEADER
WP 6	Stichting ALLAI Nederland

DISSEMINATION LEVEL	AUTHORS
Public	Catelijne Muller Monica Fernández Peñalver

Programme	Contract Number	Duration	Start
Horizon Europe	101070363	36 Months	November 1, 2022

Funded by the European Union

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

Roberta Calegari	**Alma Mater Studiorum – Università di Bologna**
Gabriel Gonzales – Castañé	**University College Cork**
Andrea Aler Tubella	**Umeå University**

Peer Reviews

Gabriel Gonzales – Castañé	**University College Cork**
Abhinay Pandya	**Kon.Philips N.V.**

Funded by
the European Union

Table of Abbreviations and Acronyms

Abbreviation	Meaning
AI	Artificial Intelligence
AI Act	Proposal for a 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 (2021)
AISIM	AI Stakeholder Identification Methodology
CEO	Chief Executive Officer
CSR	Corporate Social Responsibility
EGTAI	Ethics Guidelines for Trustworthy AI of the EU High Level Expert Group on Artificial Intelligence (2019)
EU	European Union
HLEG AI	High Level Expert Group on AI
WP	Work Package

Index of Contents

Index of Figures

Funded by
the European Union

1 Executive Summary

This deliverable serves as a preliminary overview of the vast social, legal and policy landscapes of AI-Fairness and a description of how these constantly evolving elements are identified, observed, assessed, translated, integrated, and updated.

The first aim of WP6 is to bridge the gap between the different AI-Fairness understandings and 'languages' between the technical, legal and social domains and create a common language for AI-Fairness that addresses all these dimensions. WP6 uses various methodologies to achieve this. To achieve this, an internal survey was circulated to leverage the vast and multi-disciplinary expertise around AI-Fairness within the Consortium and Advisory Board. A database has been set up to observe and analyse manifestations of AI unfairness relevant to the use case domains in society and assess the levels of awareness and development of narratives around AI unfairness.

The second aim of WP6 is to implement the social, legal and ethical notions of AI-Fairness into the AEQUITAS Engines. A preliminary overview of the ethical landscape of AI-Fairness was developed using the Ethics Guidelines of Trustworthy AI ("EGTAI") of the High-Level Expert Group on AI ("HLEG AI") of the European Commission.

The EU and European AI regulatory framework is currently being developed and has already resulted in numerous regulatory proposals, communications, strategies, declarations and self-regulatory codes that include AI-Fairness elements. Apart from that, AI never operated in a lawless world, which means that one also needs to look at existing regulations to identify legal AI-Fairness notions relevant to the use cases at hand. A mapping of vast (self-)regulatory AI-Fairness landscapes regarding the use case domains has resulted in a preliminary identification of the legal notions of AI-Fairness per use case domain.

Because AI-Fairness requires the involvement of all relevant stakeholders, an AI Stakeholder Identification Methodology ("AISIM") was developed that will become part of the Fairness-by-Design Engine.

Policies around AI are currently still evolving. The EU AI regulatory framework has not fully crystallized yet and the AI Act is still in the proposal stages, which means that the regulatory landscape will change over the duration of the project. Other developments relevant for AI-Fairness involve the AI Treaty of the Council of Europe. This treaty is currently being negotiated will determine the obligations of the Council's 47 Member States to protect human rights, democracy and the rule of law from potential harms of AI-systems. WP6 closely follows and analyses these developments to keep the project aligned with these developments.

2 Introduction

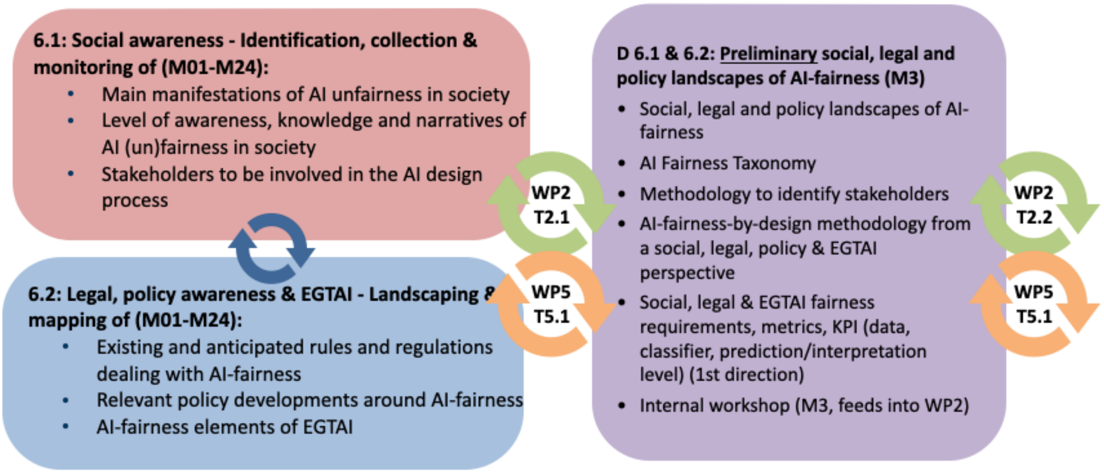
Crucial to fair and trustworthy AI, responsible innovation usually refers to anticipation, reflexivity, inclusion, and responsiveness by thinking systematically about socio-technical and legal affordances, use context, and the conflicting interests of various stakeholders. The project’s overarching methodology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

Work Package 6 (WP6) provides the AEQUITAS project with inputs needed from society to remain alert to social, legal, ethical and policy contexts and developments regarding AI-Fairness, and to ensure that these contexts and developments are incorporated in the AEQUITAS Engines: (i) Awareness and Diagnostics Engine, (ii) Reparation and Mitigation Engine, and (iii) Fairness-by-Design Engine. The social, legal, ethical and policy elements of AI-Fairness are important building blocks for the AEQUITAS Engines, that are built in WPs 3 (Awareness and Diagnostics Engine), 4 (Reparation and Mitigation Engine) and 5 (Fairness-by-Design Engine) and play a determinant role in the collection and selection of the requirements for the Engines, done in WP2. For this reason, WP6 is of an overarching nature, and constantly aligning with the other WPs. Because the topic of AI-Fairness, the social awareness and narratives around it and the ethical and legal implications of it are constantly evolving (particularly given the current regulatory and policy developments around AI, that will go through many changes in the months and years to come), the WP is also constantly adapting to the state of play. It will serve as the preliminary version of a ‘living document’ that will continuously be updated towards M24 when the final version will be presented (Deliverable 6.2).

3 Objectives and Overarching Process

One of the key objectives of WP6 is aligning the three AEQUITAS Engines with the social, legal, ethical and policy elements and contexts related to AI-Fairness. To accomplish this, the WP employs an overarching process where it interacts with other WPs on a continuous basis. As such it fully leverages the multi-disciplinary expertise and experience within the Consortium and its Advisory Board (technical, social, ethical, legal, cultural). WP6’s overarching process and interactions with the other WPs is shown in Figure 1.

Funded by
the European Union



6.1: Social awareness - Identification, collection & monitoring of (M01-M24):

- Main manifestations of AI unfairness in society
- Level of awareness, knowledge and narratives of AI (un)fairness in society
- Stakeholders to be involved in the AI design process

6.2: Legal, policy awareness & EGTAI - Landscaping & mapping of (M01-M24):

- Existing and anticipated rules and regulations dealing with AI-fairness
- Relevant policy developments around AI-fairness
- AI-fairness elements of EGTAI

D 6.1 & 6.2: Preliminary social, legal and policy landscapes of AI-fairness (M3)

- Social, legal and policy landscapes of AI-fairness
- AI Fairness Taxonomy
- Methodology to identify stakeholders
- AI-fairness-by-design methodology from a social, legal, policy & EGTAI perspective
- Social, legal & EGTAI fairness requirements, metrics, KPI (data, classifier, prediction/interpretation level) (1st direction)
- Internal workshop (M3, feeds into WP2)

Figure 1: Visual representation of interactions between WP6 and other WPs.

The technical/computer science community has been working extensively on detecting, assessing and mitigating AI *unfairness*, often by applying what we call ‘technical or statistical notions of fairness’ (such as statistical parity, predictive equality, predictive parity, etc). This approach however does not always include, or align with, ethical and legal ‘notions’ of AI-Fairness and not always considers ethical and legal specificities, requirements, and interplays when it comes to AI-Fairness. Moreover, the AI-Fairness landscape, particularly from the legal perspective, is developing quickly, with numerous AI regulations underway, resulting in stricter requirements and constraints for AI-systems when it comes to fairness. For these reasons, the AEQUITAS project aims to develop a holistic approach to AI-Fairness, considering all AI-Fairness dimensions (technical, legal, ethical and social).

The various areas of expertise and experience within the Consortium and Advisory Board bring tremendous value towards understanding AI-Fairness from multiple perspectives. The different actors involved have different understanding of what AI-fairness means and entails. The first aim of WP6 is thus to bridge the gap between these different AI-Fairness understandings and ‘languages’ and start creating a common language for AI-Fairness that incorporates all these expertise and experience.

As a first step, WP6 created an internal survey that was distributed to the Consortium and Advisory Board prior to the second AEQUITAS Consortium/Advisors meeting. The process of preparing this internal survey is visualized in Figure 2.

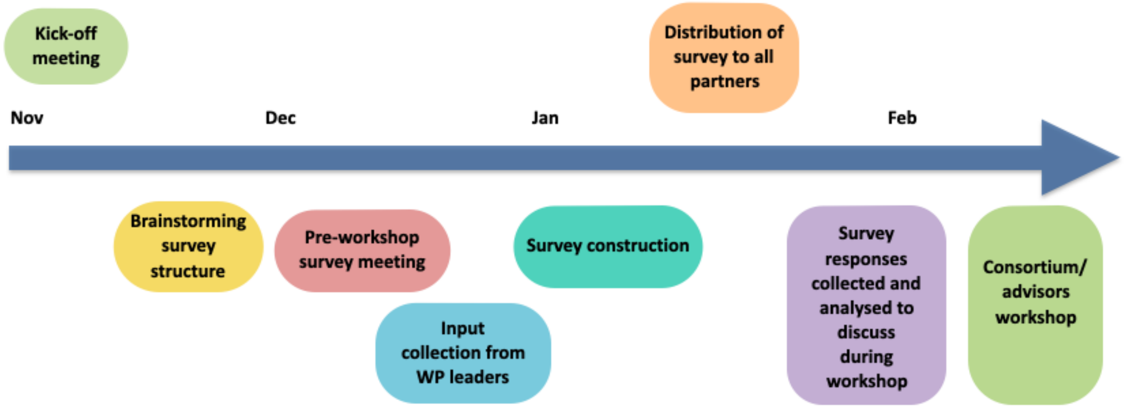

Kick-off meeting
Distribution of survey to all partners
Nov
Dec
Jan
Feb
Brainstorming survey structure
Pre-workshop survey meeting
Survey construction
Input collection from WP leaders
Survey responses collected and analysed to discuss during workshop
Consortium/ advisors workshop

Figure 2: Overview of the development process of the survey

This survey serves as a preparatory step to achieve the following objectives:

- Gain insight on the level of knowledge and awareness of technical, social, ethical and legal notions of AI-Fairness across the multiple domain expertise of the Consortium/Advisors.
- Set up the basis to construct a common AI-Fairness 'language' (Vocabulary) that incorporates/addresses all these notions.
- Set up the basis to construct a comprehensive AI-Fairness taxonomy that incorporates the technical, legal, ethical and social notions of AI-Fairness.
- Identify a preliminary set of key social, legal and technical elements regarding AI-Fairness requirements and KPIs for the AEQUITAS Engines from a social and legal perspective. This includes identifying individuals/groups at risk of being treated unfairly, social/legal/economic/cultural contexts to consider in each use case at hand.

The survey was co-created with the Engines' leaders of WP3, 4 and 5, and incorporates technical, social and legal perspectives. The survey holds the following sets of questions:

1. A set of questions on 'General AI-Fairness', to gain insights in the concerns and expectations around AI-Fairness and the level of knowledge of the various AI-Fairness dimensions (technical, social, ethical, legal) across the Consortium/Advisory Board.
2. A set of questions on the sources of AI *unfairness* (data, algorithm/model, interpretation) and ideas to assess, mitigate and avoid AI *unfairness* from the areas of expertise/experience within the Consortium and Advisory Board.
3. A set of in-depth questions on AI-Fairness in the use cases in HR, Recruiting and Candidate Selection, addressing the technical, legal and social notions of AI-Fairness, the negatively affected groups and the types of harm resulting from AI *unfairness* in these use cases.
4. A set of in-depth questions on AI-Fairness in the Healthcare use cases, addressing the technical, legal and social notions of AI-Fairness, the negatively affected groups and the types of harm resulting from AI *unfairness* in these use cases.

Funded by the European Union

5. A set of in-depth questions on AI-Fairness in the use cases regarding Child Abuse and Neglect, addressing the technical, legal and social notions of AI-Fairness, the negatively affected groups and the types of harm resulting from AI *unfairness* in this use case.
6. A set of in-depth questions on AI-Fairness in the use cases regarding the detection of Disadvantaged Students HR, addressing the technical, legal and social notions of AI-Fairness, the negatively affected groups and the types of harm resulting from AI *unfairness* in this use case.

4 Social Landscape of AI-Fairness

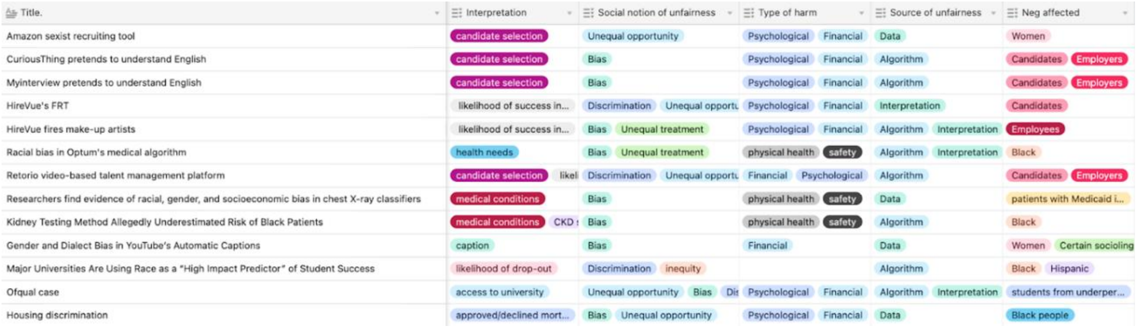
WP6 is preparing an overview of the social landscape of AI-Fairness through two processes:

1. Continuous observation and recording of the main manifestations of AI *unfairness* that (have) take(n) place in society, related to the use case domains.
2. Gathering insights in the level of awareness and knowledge of AI-Fairness across stakeholders from different domains and developing methods to include relevant stakeholders in the decision-making around, and design and development process of, AI-systems.

4.1 AI *Unfairness* Manifestations Database

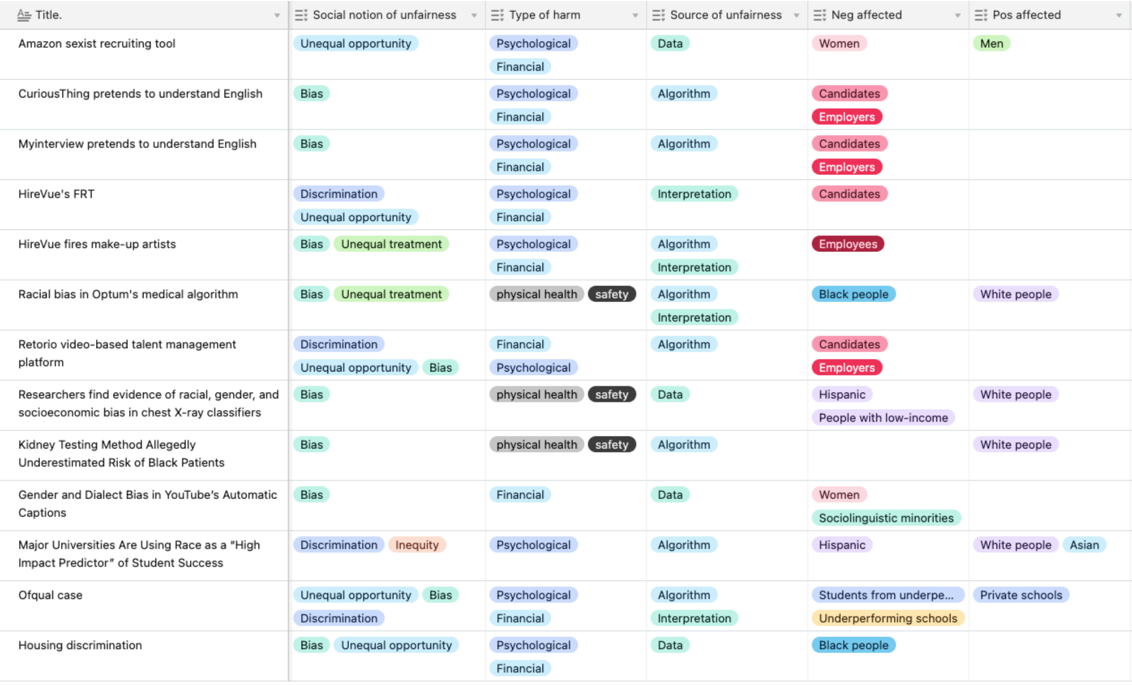
WP6 has set up a database (Figure 3 and 4) to record and analyse AI *unfairness* manifestations in society, related to the use case domains. The database allows categorization of information regarding each manifestation. The information includes:

- Technical information on the AI-technique(s), training data, input, output and interpretation.
- Source(s) of AI *unfairness* for the manifestation at hand (data, algorithm/model, interpretation).
- Ethical, legal and social notions of AI-Fairness involved.
- (Groups of) individuals negatively and positively impacted by the AI *unfairness.*
- Type(s) of harm resulting from the AI *unfairness*.
- Relevant existing/upcoming policy regarding the manifestation at hand.



Title.	Interpretation	Social notion of unfairness	Type of harm	Source of unfairness	Neg affected
Amazon sexist recruiting tool	candidate selection	Unequal opportunity	Psychological Financial	Data	Women
CuriousThing pretends to understand English	candidate selection	Bias	Psychological Financial	Algorithm	Candidates Employers
Myinterview pretends to understand English	candidate selection	Bias	Psychological Financial	Algorithm	Candidates Employers
HireVue's FRT	likelihood of success in...	Discrimination Unequal opportu	Psychological Financial	Interpretation	Candidates
HireVue fires make-up artists	likelihood of success in...	Bias Unequal treatment	Psychological Financial	Algorithm Interpretation	Employees
Racial bias in Optum's medical algorithm	health needs	Bias Unequal treatment	physical health safety	Algorithm Interpretation	Black
Retorio video-based talent management platform	candidate selection likeli	Discrimination Unequal opportu	Financial Psychological	Algorithm	Candidates Employers
Researchers find evidence of racial, gender, and socioeconomic bias in chest X-ray classifiers	medical conditions	Bias	physical health safety	Data	patients with Medicaid i...
Kidney Testing Method Allegedly Underestimated Risk of Black Patients	medical conditions CKD :	Bias	physical health safety	Algorithm	Black
Gender and Dialect Bias in YouTube's Automatic Captions	caption	Bias	Financial	Data	Women Certain socioling
Major Universities Are Using Race as a "High Impact Predictor" of Student Success	likelihood of drop-out	Discrimination inequity		Algorithm	Black Hispanic
Ofqual case	access to university	Unequal opportunity Bias Dis	Psychological Financial	Algorithm Interpretation	students from underper...
Housing discrimination	approved/declined mort...	Bias Unequal opportunity	Psychological Financial	Data	Black people

Figure 3: Screenshot of the AI-Unfairness Manifestations Database



Title.	Social notion of unfairness	Type of harm	Source of unfairness	Neg affected	Pos affected
Amazon sexist recruiting tool	Unequal opportunity	Psychological Financial	Data	Women	Men
CuriousThing pretends to understand English	Bias	Psychological Financial	Algorithm	Candidates Employers	
Myinterview pretends to understand English	Bias	Psychological Financial	Algorithm	Candidates Employers	
HireVue's FRT	Discrimination Unequal opportunity	Psychological Financial	Interpretation	Candidates	
HireVue fires make-up artists	Bias Unequal treatment	Psychological Financial	Algorithm Interpretation	Employees	
Racial bias in Optum's medical algorithm	Bias Unequal treatment	physical health safety	Algorithm Interpretation	Black people	White people
Retorio video-based talent management platform	Discrimination Unequal opportunity Bias	Financial Psychological	Algorithm	Candidates Employers	
Researchers find evidence of racial, gender, and socioeconomic bias in chest X-ray classifiers	Bias	physical health safety	Data	Hispanic People with low-income	White people
Kidney Testing Method Allegedly Underestimated Risk of Black Patients	Bias	physical health safety	Algorithm		White people
Gender and Dialect Bias in YouTube's Automatic Captions	Bias	Financial	Data	Women Sociolinguistic minorities	
Major Universities Are Using Race as a "High Impact Predictor" of Student Success	Discrimination Inequity	Psychological	Algorithm	Hispanic	White people Asian
Ofqual case	Unequal opportunity Bias Discrimination	Psychological Financial	Algorithm Interpretation	Students from underpe... Underperforming schools	Private schools
Housing discrimination	Bias Unequal opportunity	Psychological Financial	Data	Black people	

Figure 4: Screenshot of the AI Unfairness Manifestation Database

The database also allows visualization of each separate manifestation as well as the content inside each of its categories. Snapshots can for example be taken of the proportions of each type of information (e.g., type of harm, social notions of unfairness involved, techniques involved, etc.) for all or a selection of manifestations in the database (Figure 5).

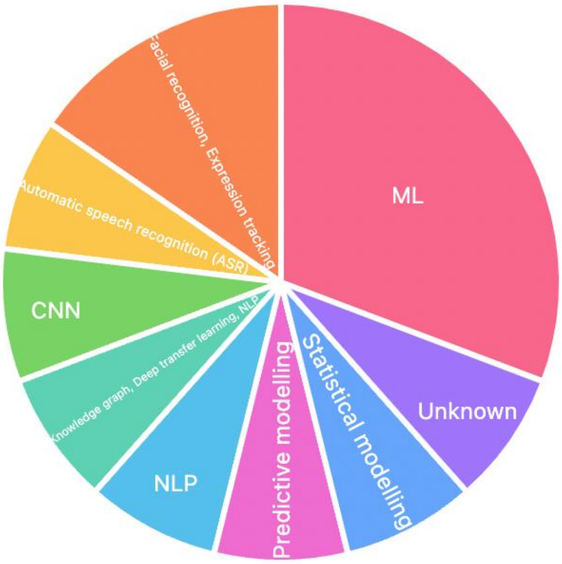


Figure 5: Example of a proportional visualization of the different AI-techniques involved in the manifestations recorded in the database.

In building the database over time, we aim to collect a large enough sample size to obtain more granular insights in the trends and distributions of AI-techniques, harms, groups affected, relevant laws, etc. over different AI *unfairness* manifestations within the various use case domains. These insights will feed into the AEQUITAS project in general and the AEQUITAS Engines in particular and ensure that the project can anticipate real-life situations related to AI *unfairness*.

4.2 AI Stakeholder Identification Methodology (AISIM)

As part of this deliverable, WP6 has 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 guide 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:

1. Stakeholders affected by the AI-system ('Affectees')
2. Stakeholders that have power over the development and deployment of the AI-system ('Decisionmakers')
3. Stakeholders that have information that would aid with the development of a fair AI-system ('Domain Experts and Users').

4.2.1 Identification of stakeholders affected by the AI-system (Affectees)

AI *unfairness* can lead to both negative and positive effects for different stakeholders depending on the case at hand. These stakeholders can include individuals or groups of individuals, such as citizens, patients, workers, students, children, parents, consumers, women, men, racial minorities, person(s) with a disability, person(s) with a low socio-economic status, person(s) with a high socio-economic status, etc. Beyond (groups of) individuals, AI-Fairness is also relevant for other types of stakeholders such as society at large, the environment, democracy, the economy, etc.

To identify stakeholders affected by AI *unfairness* the following questions serve as a guidance:

- Who/what could directly/indirectly be harmed by the AI *unfairness* in the case at hand?
- Who/what could directly/indirectly benefit from the AI *unfairness* in the case at hand?

Once identified, the stakeholders can be categorized in two groups:

- Positively affected
- Negatively affected

Per group, the stakeholders can be further categorized based on the level of impact:

- Directly affected
- Indirectly affected

Funded by the European Union

4.2.2 Identification of Stakeholders with power over the development, deployment and governance of the AI-system (Decisionmakers)

Achieving AI-Fairness requires involvement of those who have the power over the development, deployment and governance of AI-systems. These stakeholders can include a CEO or manager, a head of a Government Agency. They can include Legal Compliance officers and CSR officers. But they can also include a responsible minister or state secretary, members of parliament or a parliamentary committee, a supervisory authority, a notified body etc.

To identify stakeholders with power over the development, deployment and governance of the AI-system the following questions serve as a guidance:

- Who has the final decision to use the AI-system?
- Who is managing (aspects of) the AI project?
- Who takes care of governance of the AI-system?
- Who is auditing the AI-system?
- Who is regulating the AI-system?
- Who is supervising the AI-system?

Once identified the stakeholders can be categorized into three groups:

- Deployer (e.g., CEO, Manager, Head of Government Agency)
- Governance (e.g., Legal Compliance Officer, CSR Officer, Sustainable Development Officer)
- Authority/Supervisor (e.g., Privacy Authority, Market Authority, Financial Authority)
- Policy (e.g., Minister, Parliamentary Committee)

Per group, the stakeholders can be further categorized based on the level of involvement necessary:

- Direct involvement
- Indirect involvement
- Continuous involvement
- Ad-hoc involvement

4.2.3. Identification of Stakeholders that can aid the development of the AI-system (Domain Experts and Users)

Achieving AI-Fairness also requires involvement of those who have specific expertise regarding the domain in which the system will be used. These include technical developers such as computer scientists, machine learning experts, data scientists, statisticians. They also include domain experts with expertise on the use case at hand, such as physicians, judges, lawyers, caseworkers, teachers, financial experts, etc. But domain experts also include people who are expected to use or work with the AI-system once it has been deployed, such as salespeople, call centre employees, nurses, etc.

Funded by
the European Union

To Identify stakeholders who have information that aids the development of the AI-system the following questions serve as a guidance:

- Who is involved in the development of the AI-system?
- Who has domain expertise regarding the actions of the AI-system?
- Who (else) will be using/working with the AI-system?
- Who has a stake in understanding the workings of the AI-system?

Per group, the stakeholders can be further categorized based on the level of involvement necessary:

- Direct involvement
- Indirect involvement
- Continuous involvement
- Ad-hoc involvement

5 Ethical and Legal Landscapes of AI-Fairness

WP6 works on identifying the vast ethical and legal landscapes of AI-Fairness related to the use cases and has prepared a preliminary overview of these landscapes.

5.1 Preliminary overview of the 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. WP6 analysed EGTAI and found that elements of AI-Fairness are engrained throughout it. First and foremost, EGTAI identifies ‘Fairness’ as one of the four main ethical principles relevant for AI, along with the principles of ‘Respect for Human Autonomy’, ‘Prevention of Harm’ and ‘Explicability’. Consequently, elements of fairness can be found in the 7 key requirements for Trustworthy AI of EGTAI, endorsed by the European Commission.

5.1.1 Fairness related to Key requirement 1: Human Agency and Oversight

Ensuring human agency and oversight is important for AI-Fairness because they ensure that individuals are not unfairly manipulated, deceived, herded or conditioned, or subject to any other unfair outcomes. The former ensures that human autonomy regarding decision making is respected. Human oversight requires that systems are regularly reviewed and audited to ensure they are operating as intended. Together, they help address and prevent any issues of unfairness (such as bias or discrimination) coming from the AI-system.

Funded by
the European Union

5.1.2 Fairness related to Key requirement 2: Technical Robustness and Safety

Ensuring technical robustness and safety of AI-systems is important for AI-Fairness because it helps to prevent errors or unintended consequences that could result in bias, discrimination and other instances of unfair treatment. This includes ensuring that AI-systems are reliable, accurate, secure, resilient to attacks, and have a fallback plan. Together these requirements ensure that there are no unexpected outcomes or attacks that may lead to harm or unfair results.

5.1.3 Fairness related to Key requirement 3: Privacy and Data Governance

Ensuring privacy and data governance is important for AI-Fairness because it helps protect individuals' personal information and prevent discrimination based on sensitive characteristics such as race, gender, or sexual orientation. This means that 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.

5.1.4 Fairness related to Key requirement 4: Transparency

Ensuring transparency is important for AI-Fairness because it 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*.

5.1.5 Fairness related to Key requirement 5: Diversity, Non-discrimination and Fairness

Diversity, inclusion, and non-discrimination in the design of an AI-system helps ensure that the system is fair. These principles not only aim to 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. Lastly, 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, can ensure that an AI-system does not produce any unfair outcomes towards certain groups.

5.1.6 Fairness related to Key requirement 6: Social and Environmental Well-being

Social and environmental wellbeing is important for AI-Fairness because it 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. All which are elements of a fair society.

5.1.7 Fairness related to Key requirement 7: Accountability

Accountability is closely linked to the principle of fairness and relevant to AI-Fairness because it helps to ensure that individuals and organizations can be held accountable for any issues of bias or discrimination that may arise from the use of AI-systems. This includes establishing clear lines of responsibility and liability, minimizing and reporting negative impacts and trade-offs that influence fairness.

This preliminary landscape has led us to the conclusion that the Ethics Guidelines of Trustworthy AI are well suited to aid in the creation of white list rules and black list restrictions for AI that are linked to the principle of fairness.

5.2 Preliminary Overview of the Legal Landscape of AI-Fairness

The EU and European AI regulatory framework is currently being developed and has already resulted in numerous regulatory proposals, communications, strategies, declarations and self-regulatory codes that include AI-Fairness elements. Apart from that, AI never operated in a lawless world, which means that one also needs to look at existing regulations to identify legal AI-Fairness notions relevant to the use cases at hand.

In principle, three (self-)regulatory dimensions are relevant for AI-Fairness in HR, Recruiting and Candidate Selection:

1. European Union dimension
2. EU Member States’ dimension
3. European dimension

Within each dimension, multiple (self-)regulatory instruments (treaties, regulations, directives, laws, ethics guidelines, social partner agreements and other instruments) exist that deal with elements of fairness in the domain at hand. These instruments have different implications for different actors (states, private actors, citizens, etc.) and often interact or influence each other. Often there exists a ‘cascading’ effect, where a regulatory instrument is set at EU or European level, and then applied to all Member States. This cascading effect can be fully harmonizing, where the application should be equal in all Member States, or directional, where the regulatory instrument merely serves as a minimum standard and states can set their own rules regarding the topic.

WP6 has initiated a mapping of these vast (self-)regulatory AI-Fairness landscapes regarding the use case domains, which has resulted in a preliminary identification of the (self-)regulatory AI-Fairness notions relevant for each use case domain. This mapping exercise resulted in a preliminary overview of (generalized) legal and self-regulatory notions of AI-Fairness per use case domain.

Funded by
the European Union

5.2.1 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. WP6 preliminary identified the following notions in existing instruments: non-discrimination; right to engage in work; freedom to choose an occupation; freedom of religion; right to private and family life; freedom of movement of workers; gender equality; racial equality; equal employment; data protection; positive discrimination; equal treatment; equal opportunity; avoidance of harassment; fair, transparent and explainable digital HR and recruiting; vigilance of the risk of compromising human dignity in digital HR and recruiting; human oversight in digital HR and recruiting; respect for GDPR; right to human intervention in digital HR and recruiting.

The proposal for the 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. Once the AI act comes into force, these systems will have to comply with a large set of requirements before they can be put on the EU internal market. The requirements are aimed at protecting health, safety and fundamental rights, from harmful effects of AI in this domain. The 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 and (v) accuracy, robustness and cybersecurity. These ‘overarching’ categories stem from EGTAI and are all (same as the 7 key requirements of EGAI) relevant for AI-Fairness.

Moreover, the proposal for the AI Act prohibits social scoring by public authorities, which it describes as (paraphrased) 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.

5.2.2 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.

WP6 preliminarily identified the following notions in existing regulatory instruments: 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 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 as high risk. Separately, the proposal for the AI Act classifies biometric identification and categorization as high risk. Once the AI Act comes into force, these systems must comply with a large set of requirements before they can be put on the EU internal market. The requirements are aimed at protecting health, safety and fundamental rights, from harmful effects of AI in this domain. The 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 and (v) accuracy, robustness and cybersecurity. These ‘overarching’ categories stem from EGTAI and are all (same as the 7 key requirements of EGAI) relevant for AI-Fairness.

5.2.3 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.

For the use case regarding *detection of child neglect and abuse*, WP6 preliminary identified the following fairness notions in existing regulatory instruments: 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.

The proposal for the AI Act classifies biometric identification and categorization as high risk. Separately, the proposal for the AI Act classifies biometric identification and categorization as high risk. Once the AI Act comes into force, these systems must comply with a large set of requirements before they can be put on the EU internal market. The requirements are aimed at protecting health, safety and fundamental rights, from harmful effects of AI in this domain. The 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 and (v) accuracy, robustness, and cybersecurity. These ‘overarching’ categories stem from EGTAI and are all (same as the 7 key requirements of EGAI) relevant for AI-Fairness.

.

Funded by
the European Union

For the use case on *access to education for disadvantaged students*, WP6 preliminary identified the following fairness notions in existing regulatory instruments: 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 proposal for 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. Separately, the proposal for the AI Act classifies biometric identification and categorization as high risk. Once the AI Act is adopted, these systems must comply with a large set of requirements before they can be put on the EU internal market. The requirements are aimed at protecting health, safety and fundamental rights, from harmful effects of AI in this domain. The 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 and (v) accuracy, robustness, and cybersecurity. These 'overarching' categories stem from EGTAI and are all (same as the 7 key requirements of EGTAI) relevant for AI-Fairness.

6 Policy Developments around AI-Fairness

Continuously following policy developments around AI is crucial, as these are currently evolving. The EU AI regulatory framework has not fully crystallized yet and the AI Act is still in the proposal stages, which means that the text of the AI Act will change over the duration of the project. Other developments relevant for the use cases involve the AI Treaty of the Council of Europe. This treaty is currently being negotiated will determine the obligations of the Council's 47 Member States to protect human rights, democracy and the rule of law from potential harms of AI-systems. It will hold numerous elements relevant for AI-Fairness and is set to be finalized by the end of 2023. The joint EU trade unions have called for an EU Directive on Algorithmic Systems at Work, which might be relevant for the AEQUITAS use case in HR, recruiting and candidate selection. The European Commission proposed a European Health Data Space Act, relevant for the use cases in the healthcare domain.

WP6 is scanning, observing and analysing these developments on a continuous basis, through various methods, such as:

- Direct exchanges with European and national policy makers on AI policy and regulation.
- Review and analysis of policy positions from the European Council and European Parliament on the upcoming AI Act and AI-related regulatory proposals at EU level.

Funded by
the European Union

- Active participation in and organization of exchanges with national parliaments and governments on AI policy and AI regulation.
- Active participation in negotiations and deliberations on the AI Treaty text between the member states of the Council of Europe.
- Review and analysis of the draft negotiating texts of the AI Treaty.
- Research on the implications of AI policy and regulatory developments for AI-Fairness.

7 AI Fairness-by-Design Methodology

WP6 will develop an AI Fairness-by-Design Methodology that covers the entire lifecycle of an AI-system comprising of the following phases:

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

For the scoping phase (1), WP6 constructed the following 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.

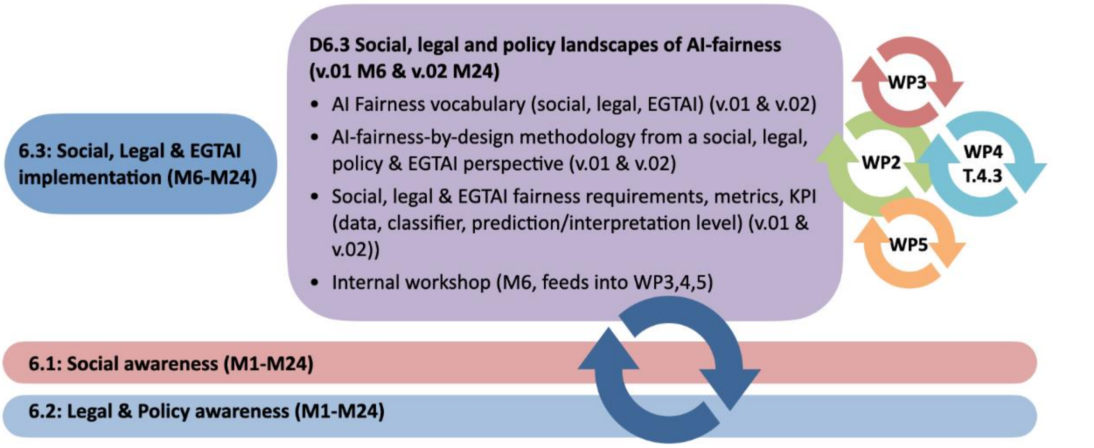
8 Conclusions and next steps

Through this deliverable, WP6 provides the AEQUITAS project with inputs needed from society to remain alert to social, legal, ethical and policy contexts and developments regarding AI-Fairness. These inputs cover the social, legal, and ethical landscapes of AI-Fairness and methodologies that can be integrated into the AI Fairness-by-Design methodology.

The elements of AI-Fairness identified and observed during the completion of this deliverable are important building blocks for the AEQUITAS Engines built in WPs 3, 4 and 5 and play a determinant role in the collection and selection of the requirements for the Engines, done in WP2. Thus, in the next steps, WP6 will collaborate with WP's 2, 3, 4, and 5 in implementing these social, legal and EGTAI notions of fairness into the

AEQUITAS' Engines. WP6 will further build on the preliminary social, ethical and legal landscapes by:

- Continuing to identify and analyse manifestations of AI *unfairness* in the use cases' domains and further build the database.
- Developing an adapted survey considering the outcomes of the internal discussions, and circulated among focus groups, composed of members of the constituency of the stakeholder partners within the Consortium (Period Think Tank, Acrigay, EUROCADRES, Women in AI, Asociacion Rayuela).
- Analysing and assessing the EU and European regulatory and self-regulatory instruments identified in the preliminary legal landscapes and further identifying the relevant AI-Fairness elements.
- Co-determining AI-Fairness requirements, metrics and KPI for the Engines derived from the social, legal and ethical notions of fairness.
- Further developing the AI Fairness-by-Design methodology.
- Continuously observing, analysing and integrating relevant policy developments around AI-Fairness.
- Mapping the involvement of relevant stakeholders against the AI-lifecycle using AISIM.



D6.3 Social, legal and policy landscapes of AI-fairness (v.01 M6 & v.02 M24)

- AI Fairness vocabulary (social, legal, EGTAI) (v.01 & v.02)
- AI-fairness-by-design methodology from a social, legal, policy & EGTAI perspective (v.01 & v.02)
- Social, legal & EGTAI fairness requirements, metrics, KPI (data, classifier, prediction/interpretation level) (v.01 & v.02))
- Internal workshop (M6, feeds into WP3,4,5)

6.3: Social, Legal & EGTAI implementation (M6-M24)
WP3
WP2
WP4
T.4.3
WP5
6.1: Social awareness (M1-M24)
6.2: Legal & Policy awareness (M1-M24)

Figure 6: Overview of the next steps towards implementation of social, legal and ethical notions of-Fairness

AEQUITAS
unbias AI

Consortium

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