AEQUITAS

unbias AI

Interactive Case Studies

Consortium

UMEÅ UNIVERSITET

UCC University College Cork, Ireland / Coláiste na hOllscoile Corcaigh

THE ADECCO GROUP

AKKODIS

SERVIZIO SANITARIO REGIONALE EMILIA-ROMAGNA / POLICLINICO DI SANT'ORSOLA

PHILIPS

LOBA®

ALLAI.

PERIOD think tank

ARCIGAY Associazione LGBTI+ Italiana

EUROCADRES

ITI INVESTIGATE TO INNOVATE

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Universidad de La Laguna

AR Asociación Rayuela

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

Fair-by-Design Building Blocks: AEQUITAS developed tools and methodologies to translate the user requirements to "building blocks", which align with the AI lifecycle to facilitate the usage of AEQUITAS tool by the AI developers or other users. The building blocks include the following methodology components: Stakeholder Identification Methodology, Stakeholder Engagement Methodology, AI Fairness Readiness Assessment, Fundamental Rights Impact Assessment for Fairness, AI Act Risk Classification, Social, Legal, and Ethical Fairness Constraints Identification, Fair Data Collection, Governance, and Management Methodology, Fair Model Selection/Design, Addressing Interpretation Bias. All building blocks are interconnected. As the project progresses, the placement of these building blocks might be adjusted. For example, Social, Legal, and Ethical Fairness Constraints Identification could be expanded and moved to the scoping phase. Further details on these methodologies can be found in D6.3.

The QA component within AEQUITAS is designed to guide decision-makers toward creating fairer AI systems. The questions and answers are developed by experts in sociology, law, statistics, and computer science. The main actors include Business User (BU), Technical User (TU), End User, Stakeholder, and Experts. The process begins with initial questions aimed at assessing the presence or absence of relevant datasets or algorithms. The sequence of questions is dynamic, depending on the starting point and the answers provided up to that point. The Technical User (TU) may leverage automated techniques to provide suggestions and recommendations to the Business User (BU). The system addresses the need for a tool that assists organizations in creating AI systems from various starting points, whether they have datasets, pre-trained algorithms, both, or neither. The system employs a question-answering (QA) approach to guide the BU's decision-making process. It allows the TU to provide necessary data, reports, summaries, and implementations, thereby facilitating coordination between the BU and TU.

For instance, in the ADECCO use case (HR1), a dataset of job applicants is used to automate the selection process via AI, ensuring fairness. A BU from ADECCO initiates a new project using the AEQUITAS frontend. The BU answers a series of questions to define fairness requirements and constraints, identify sensitive features and protected groups, and select fairness notions, metrics, and mitigation algorithms. After this stage, the BU transitions the task to the decision support system (DS), which performs fairness measurements and mitigations via the Core Library. When controlling data distributions is necessary, a synthetic data generator is utilized. With this awareness-raising tool, AEQUITAS allows users to explore bias in decision-making and analyse the outcomes of both biased and debiased systems. Further details can be found in D5.4.

The recruitment case study was selected to gather inspiration, draw conclusions for the proposed bias detection approach, and validate its effectiveness. The AEQUITAS recruitment case study involves a complex system comprising three primary components: Work Order Generation System identifies client needs. Candidate and Job Posting Management System features matching capabilities and integrates with external systems. Application Management Platform manages candidate applications. The system collects various candidate features, including personal Information, name, surname, email, phone number, professional details, and location data. Stakeholders have expressed interest in implementing AI solutions to enhance these processes. Their primary focus is on developing an AI-assisted recruiting tool that reduces gender and other biases while replacing human oversight in the current system. Specific goals include minimizing cognitive and structural biases related to age, race, gender, education, socioeconomic background, location, disabilities, language and communication styles, career gaps and non-traditional career paths. In addition to addressing these biases, stakeholders aim to avoid predictive bias, track potential candidates who haven't applied, identify the most suitable candidates based on specific criteria, predict candidate performance and employee retention, and recommend training and career development opportunities.

The case study presents and compares two approaches: the Traditional Direct Matching Flow and an AI-Enhanced Alternative Flow. These use cases are integrated into educational and awareness-raising tools to address the social dimensions of AI fairness. Further details can be found in D3.3.

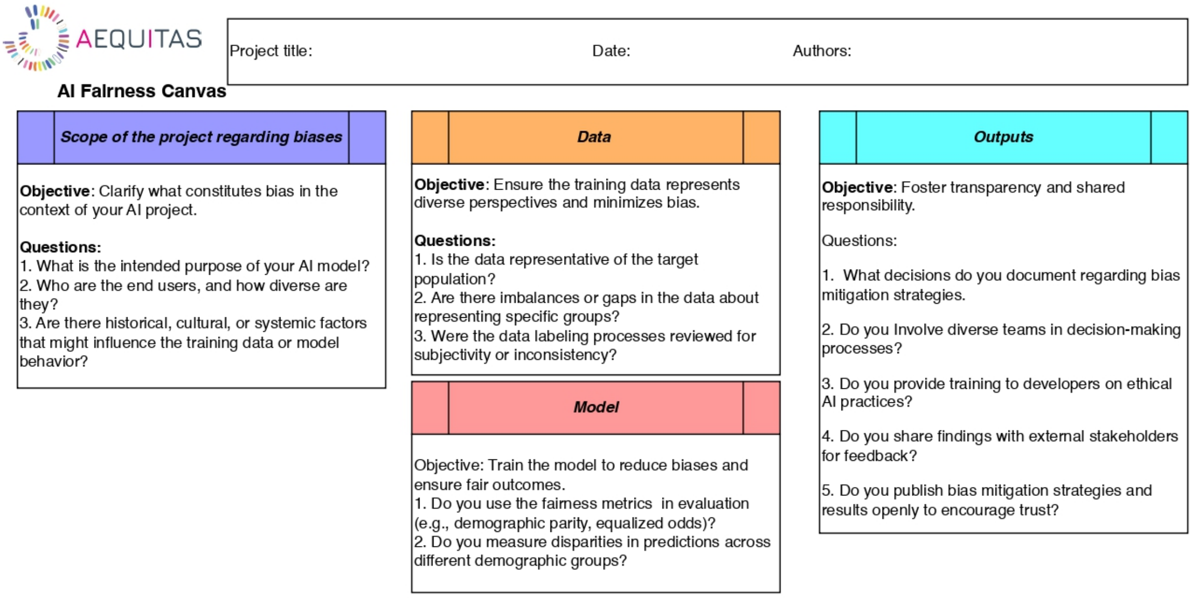
Role playing in case study: Participants are expected to form teams to analyse and discuss a case study. The objective of this assignment is to help you apply the lessons learned throughout the course to a practical scenario. The expected outcome is to enable educators to better understand the complexities of AI systems and the need to assess fairness in their design and implementation. The case study, "Hiring by Machine", was developed by the University Center for Human Values (UCHV) and the Center for Information Technology Policy (CITP) at Princeton. For this exercise, the focus has been narrowed to discussions and reflections on fairness from social and legal perspectives.

Role-Playing Activity: Each group member will assume one of the following roles:

- Hiring manager
- HR Interviewer
- Technique Interviewer
- Data scientist who (is responsible) develops this AI recruitment system
- Applicant (interviewee)

More details on the background and specific instructions for this activity are available in D6.5. Each group is required to submit an integrated report summarizing their reflections and conclusions based on the outlined steps.

An AI Fairness Canvas as shown in Figure 3 was developed as an interactive exercise to be included in educational materials. This exercise is aimed at AI developers and researchers, encouraging them to explore bias-related questions and identify actions to address these issues in their projects.



AEQUITAS

AI Fairness Canvas

Project title: Date: Authors:

Scope of the project regarding biases

Objective: Clarify what constitutes bias in the context of your AI project.

Questions:
1. What is the intended purpose of your AI model?
2. Who are the end users, and how diverse are they?
3. Are there historical, cultural, or systemic factors that might influence the training data or model behavior?

Data

Objective: Ensure the training data represents diverse perspectives and minimizes bias.

Questions:
1. Is the data representative of the target population?
2. Are there imbalances or gaps in the data about representing specific groups?
3. Were the data labeling processes reviewed for subjectivity or inconsistency?

Model

Objective: Train the model to reduce biases and ensure fair outcomes.
1. Do you use the fairness metrics in evaluation (e.g., demographic parity, equalized odds)?
2. Do you measure disparities in predictions across different demographic groups?

Outputs

Objective: Foster transparency and shared responsibility.

Questions:

1. What decisions do you document regarding bias mitigation strategies.

2. Do you Involve diverse teams in decision-making processes?

3. Do you provide training to developers on ethical AI practices?

4. Do you share findings with external stakeholders for feedback?

5. Do you publish bias mitigation strategies and results openly to encourage trust?

Figure 1 AEQUITAS canvas – an exercise in lecture on AI fairness to AI developers/researchers

3.3 Guidelines and Resources

Different countries and regions have implemented regulations and laws aimed at promoting trustworthy AI. In Europe, the 2019 Ethics Guidelines for Trustworthy AI, established by the High-Level Expert Group on AI (HLEG AI), define trustworthy AI as systems that are Lawful, Ethical, and Robust. These principles are further elaborated through four key trustworthiness pillars including Respect for human autonomy, Prevention of harm, Fairness, and Explicability. To operationalize these principles, the guidelines outline seven concrete requirements for achieving trustworthy AI, as depicted in Figure 4. Furthermore, the AI Act marks a significant regulatory milestone, strengthening the EU’s commitment to developing and deploying AI systems responsibly and in alignment with societal values and regulatory standards. The AEQUITAS project, especially WP6, has investigated numerous regulatory frameworks and guidelines to ensure the ethical and lawful use of AI.

Key initiatives include the Ethics Guidelines for Trustworthy AI and new AI-related legislation such as the Digital Services Act, Digital Markets Act, Transparency and Targeting of Political Advertising Act, Platform Worker Directive, Data Act, Data Governance Act, European Health Data Space Act, and the AI Liability Directive. The AEQUITAS project investigates both primary fairness regulations, such as the EU Charter of Fundamental Rights, which addresses principles like Human Dignity (e.g., life, physical and mental integrity, privacy, fair trial, reasonable suspicion, and rights of the child), Freedom (e.g., expression, information, assembly), Fairness (e.g., non-discrimination, equal treatment of men and women, elderly individuals, and people with disabilities), Social Rights (e.g., education, right to work, and freedom to choose an occupation). It also examines secondary fairness regulations, such as Racial Equality Directive, Gender Equality Directive, Unfair Commercial Practices Directive, Freedom of

Movement of Workers Regulation, Equal Employment Directive, General Data Protection Regulation (GDPR), Gender Access Directive.

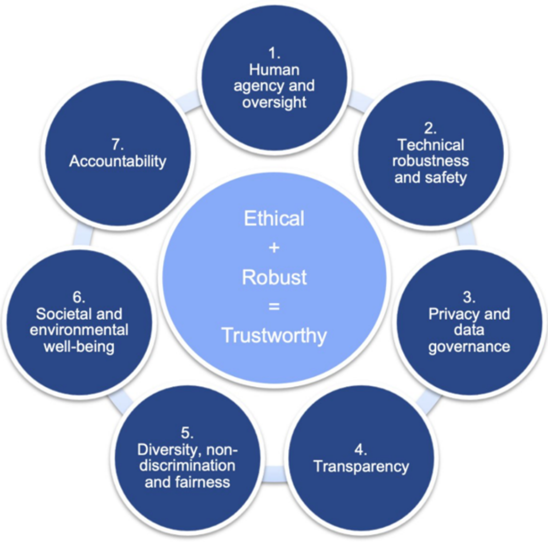

1.
Human agency and oversight
2.
Technical robustness and safety
3.
Privacy and data governance
4.
Transparency
5.
Diversity, non-discrimination and fairness
6.
Societal and environmental well-being
7.
Accountability
Ethical
+
Robust
=
Trustworthy

Figure 2 Seven trustworthy requirements of AI systems (EU)

It is essential to bridge the gap between the technical, legal, and social domains of AI fairness. AEQUITAS applies various methodologies to integrate social, legal, and ethical notions of AI fairness into its engines. The AEQUITAS project provides guidelines to help society remain vigilant about the social, legal, ethical, and policy contexts surrounding AI fairness, as detailed in D6.1. A preliminary ethical landscape was developed using the Ethics Guidelines for Trustworthy AI (EGTAI) proposed by the High-Level Expert Group on AI (HLEG AI) of the European Commission. The social and legal landscape of AI fairness is explored to provide a checklist (as outlined in D6.3). Through it, AEQUITAS offers a preliminary overview of possible social and legal methodologies to address fairness in the design of AI systems at the data, classifier, and prediction levels.