AI Fairness-by-Design Multi-Stakeholder Methodology

A Comprehensive Framework for Fair AI Design and Development

AEQUITAS
unbias AI

A Multistakeholder Approach to AI Fairness-by-Design (MAP) - Meta Methodology

Abbreviation	Meaning
AFF	Affectees
AIU	AI Users
DDM	Development Decisionmakers
DE	Domain Experts
EGTAI	Ethics Guidelines for Trustworthy AI
FbD	Fair-By-Design
FDCGM	Fair Data Collection, Governance, and Management
FMM	Fair Model Methodology
FOIM	Fair Output Interpretation Methodology
FRIA-F	Fundamental Rights Impact Assessment for Fairness
GDM	Governance Decisionmakers
SIM	Stakeholder Identification Methodology
TAIRA	Trustworthy AI Readiness Assessment
MAP	Multistakeholder Approach to AI Fairness-by-Design
ML	Machine learning
NLP	Natural Language Processing

AEQUITAS
unbias AI

Contents

Introduction

Fair AI systems need a collaborative socio-technical approach that helps us: a) understand the potential legal, ethical and social effects of the AI system and improve the design and implementation choices based on that understanding; b) audit algorithms and their output to make any biases transparent; and c) continuously monitor the workings of the systems to mitigate the ill effects of any biases.

This meta-methodology aims to provide a clear roadmap on the role of AI stakeholders, their continued engagement process, their various engagement formats at any given moment during the **scoping, risk analysis,** and **development stages** of the AI lifecycle, and the sub-methodologies they can utilize to execute their engagement.

Stakeholders and Engagement

Identification of Stakeholders

Stakeholder engagement starts with identifying which stakeholders to engage. The AI Stakeholder Identification Methodology (AI SIM), which is a process developed by AEQUITAS WP6, guides the identification of all relevant AI stakeholders for any given AI design project. To make the methodology product and domain agnostic, it builds on three stakeholder groups: (i) Affectees (AFF), (ii) Decisionmakers and (iii) Domain Experts (DE) & Users. A set of questions helps identify specific actors within each stakeholder group. Furthermore, the type and level of involvement of each stakeholder is identified. The table below summarises the AI SIM for a better overview of the methodology.

AI Stakeholder Groups	How to identify them?	1st categorisation	Levels of Involvement
Affectees (AFF): Stakeholders affected by the AI-system	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?	Positively affected Negatively affected Directly affected Indirectly affected	Direct involvement Indirect involvement Continuous involvement Ad-hoc involvement
Decisionmakers: Stakeholders that have power over the development and deployment of the AI-system	Who is involved in the development of the AI-system? Who is managing (aspects of) the AI project?	**Development Decisionmakers (DDM):** e.g. AI developers, data scientists, AI providers[1] **Governance Decisionmakers (GDM):** (socio-legal and ethical expertise, management):	Direct involvement Indirect involvement Continuous involvement Ad-hoc involvement

[1] 'provider' means a 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 ((3), Article 3, the EU AI Act).

	Who has the final decision to use the AI-system? Who takes care of governance of the AI-system? Who is auditing the AI-system? Who is supervising the AI-system?	e.g. Legal Compliance Officer, CSR Officer, Sustainable Development Officer, Manager, CEO **Authority/Supervisor:** e.g. AI Act authority, Privacy Authority, Market Authority, Notified Body, AI Office	
Domain Experts (DE) and AI Users (AIU): Stakeholders that have information that would aid with the development of a fair 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?	**Domain Experts (DE)** with expertise on the use case at hand **AI deployers**[2] that do not develop the system themselves but procure and deploy it **AI Users (AIU)** that will work with the AI system in their day-to-day[3]	Direct involvement Indirect involvement Continuous involvement Ad-hoc involvement

Table 1: Summary of the AI SIM

All three AI stakeholder groups may be internal or external to the organisation, depending on its size and the type of skills it has available. Additionally, another typology that can be considered is the stakeholder's specific background or expertise.

Delimitation of the Stakeholder Engagement and Role Typology

To ensure Fair-by-Design AI systems, the identified stakeholders should collectively contribute to and collaborate throughout the AI lifecycle stages. The process must incorporate the collaboration of interdisciplinary teams, as already expressed in Deliverable 5.1, where their engagement and collaboration need to be organized in a harmonized flow of roles for an efficient AI-Fairness establishment. Additionally, AI bias is a risk that should be eliminated or sufficiently mitigated, the absence of which could result in legal liability for damage or penalties. Therefore, AI stakeholders' roles should be clearly defined for accountability and transparency purposes.

The necessity for a clear role delimitation of each stakeholder (group) in the process involves three types of engagement:

- Responsible (leading or guiding the process)
- Co-responsible (actively assisting responsible stakeholder)
- Provider of Domain Expertise or Feedback

[2] '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 a personal non-professional activity ((4), Article 3, the EU AI Act).

To identify the role of each stakeholder in the FbD methodology, throughout the AI lifecycle stages, the following set of questions may be of assistance:

- How will **Governance Decisionmakers (GDM)** guide the execution of the FbD building blocks, throughout the AI lifecycle stages, jointly with **Development Decisionmakers (DDM)**?
- How will the **Governance Decisionmakers** ensure all other relevant stakeholders are included and engaged within the process?
- How will the **Joint Decisionmakers** (the Governance Decisionmakers and Development Decisionmakers) along with **Domain Experts (DE) and AI Users (AIU)** (All groups) and **Affectees (AFF)** collectively concretise a Fair AI System?

The responsible Stakeholder during an AI lifecycle stage is the entity that will be leading the steps to be taken during that stage and making sure AI-fairness requirements are implemented.

In the stages where interdisciplinary, cross-sectorial, or any other type of collaboration is needed, clarification of who will be the **Co-responsible Stakeholder** will be highlighted in this document and specified under each AI lifecycle stage. The collaboration between the responsible and co-responsible stakeholders could be general during an AI lifecycle stage, or specific, as regards the execution of a particular sub-methodology.

The role of **expertise and feedback Providers** is consultative and reflective. The aim of this role is to enrich the process with concrete feedback on AI fairness and get experts' insights on use case specificities. The FbD methodology will follow a set of participatory methods to engage stakeholders that are usually not at the table of decision such as affectees (AFF), domain experts (DE), and AI users (AIU).

For concretisation, the table below is a codified illustration of the type of AI stakeholders' engagement in each AI lifecycle stage. A detailed analysis of each stakeholder's function and tasks will be provided and personalized for each building block.

AI stakeholders	**Scoping**	**Risk**	**Development**	**Evaluation**
Affectees **(AFF)**	▲	▲		▲
Governance Decision Makers **(GDM)**	◆	◆	◆	◆
Development Decision Makers **(DDM)**	◆	◆	◆	◆
Supervisors Notified Bodies				◆
Domain Experts **(DE)**	▲	▲		▲

AI Users **(AIU)**	▲	▲		▲

◆ Responsible
◆ Co-responsible
▲ Sharing expertise/feedback

Table 2: Stakeholder function overview

The **Governance Decision Makers (GDM)** are the group with decision-making power over the development and deployment of the AI-system, and should include ethical, legal and social expertise. They will be the **Responsible stakeholder** throughout the AI lifecycle stages in leading and guiding the accomplishment of all FbD building blocks (sub-methodologies). They will have an overarching responsibility throughout the process, ensuring a harmonized and efficient execution of and transition from each task while enabling the enforcement of interdisciplinarity to sustain the comprehensive approach AEQUITAS promotes.

The **Development Decision Makers (DDM)** are the stakeholders that have power over the development and deployment of the AI-system from a technical perspective and that are usually the developers of the technology. This group will be the **Co-responsible** stakeholder who will actively assist the GDM throughout the process with their technical expertise and execute the tasks that are directly related to their field.

The **Affectees (AFF)**, **AI Users (AIU)**, **Domain Experts (DE)**, **Fairness Feedback (FF)** providers, and **Supervisors/Notified Bodies**, are stakeholders that should be included at various stages of the process. They will take different engagement roles as the table above illustrates. Depending on the AI lifecycle stage, some will be **Co-responsible**, and others will enrich the FbD steps with **their expertise and feedback**. The details of their engagement roles and tasks will be further explained in this document under each AI lifecycle stage.

This meta-methodology aims to be overall legally, ethically and socially agnostic serving AI stakeholders worldwide. However, in this document, EU frameworks such as the AI Act and the EU Charter of Fundamental Rights and the EU Ethics Guidelines for Trustworthy AI are incorporated in the building blocks to illustrate how regulations, policies, or frameworks may be included within the AI lifecycle stages.

A co-creative Engagement Format

AI Stakeholders will engage in a **co-creative process** in the FbD methodology. This is a collaborative approach where stakeholders work together on a shared task or project, often in a physical or virtual space. This approach emphasizes hands-on participation, teamwork, and co-building, enabling participants to contribute actively to the design, development, deployment, and use of AI systems. It fosters a sense of ownership and collective responsibility for the outcomes, enhancing engagement, and incorporating diverse perspectives.

Between the **responsible and co-responsible stakeholders**, collaboration will be through workshops, meetings, and contributions to deliverables and documentation. The **engagement of expertise and feedback providers** will follow a structured set of participatory methods to

enable representativity and accuracy of their engagement formats in regard to the aim of their contribution.

Participatory methods have the potential to address **issues of power** and **inclusion** in AI (as noted in deliverable 5.1), but their benefits and challenges in practice are still unclear because few organizations have deeply engaged with them when it comes to AI. In this section, we explain how stakeholders (expertise and feedback providers) will be sharing their inputs through a set of participatory methods including actors that are not traditionally incorporated in the process, to enable inclusivity and representation.

Below is a list of **participatory methods** and their definitions to explain the considered ways of participation in the SEM. These formats will be highlighted in each lifecycle stage of the AI system later in this deliverable.

🔎 **Focus Group:** Focus groups are small, diverse, and representative groups of people brought together to facilitate in-depth discussions and practical work, allowing stakeholders to provide detailed feedback and insights. This format is valuable for gathering qualitative data and exploring participants' perceptions, opinions, and attitudes in a collaborative setting.

📋 **Survey:** Standardized questionnaires to collect data about people and their preferences, thoughts, and behaviours systematically. These can gain structured feedback from AI Users (AIU) and Affectees (AFF). They allow for gathering quantitative data and insights on users' and Affectee's (AFF) experiences, needs, and concerns, which can be analysed to inform decision-making and improve AI systems.

💬 **Roundtable:** Roundtables are a form of simplified discussion platform which enables stakeholders to discuss issues in an open environment. This technique is suitable to brainstorm, collect feedback, and overview stakeholders' opinion and practices.

The Multi-Stakeholder Approach per AI lifecycle Phase

In this chapter, we will describe the methodology of stakeholder engagement during the different AI lifecycle phases and identify the sub-methodologies to be employed at each phase.

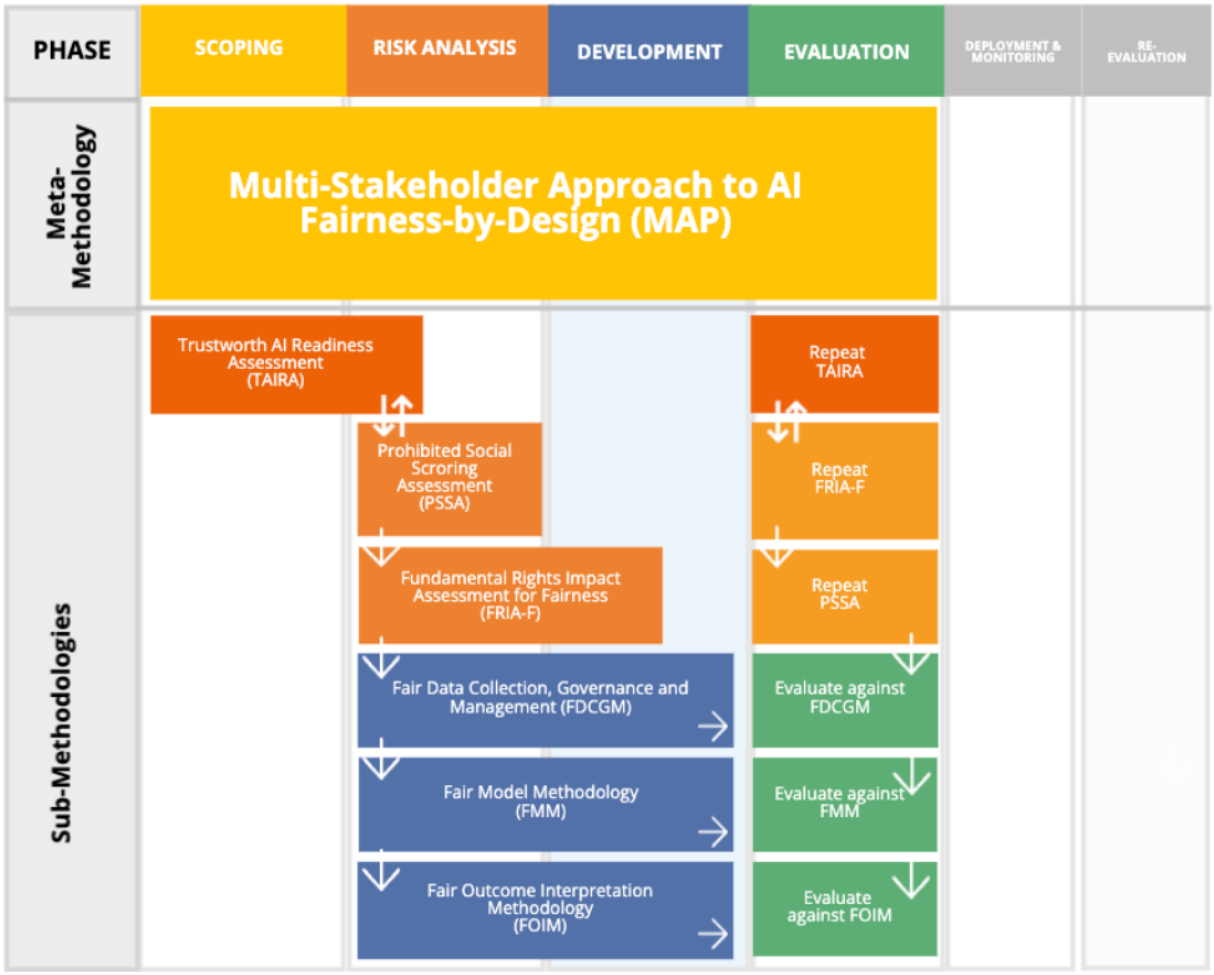

Figure 2. Visualisation of the various sub-methodologies mapped against the AI lifecycle stages.

Scoping Phase

Scoping is the first phase of the AI lifecycle where the framework guiding the Fair-by-Design methodology is established. It includes a Stakeholder Identification, and execution of the AI Fairness Readiness Assessment (Annex I).

Goals during the Scoping Phase:

📍 Establishing necessity, proportionality and fairness readiness
📍 Identifying stakeholders and planning stakeholder engagement

Engaged Stakeholders during this Phase:

🔷 Responsible stakeholders: **Governance Decisionmakers** (legal, ethical experts and managers) as they guide the process towards and within the scoping phase. They should make sure that all building blocks are completed, facilitate the transit from one task to another, and coordinate between the engaged stakeholders. **GDM** will conclude the scoping stage by documenting the results in a '**Trustworthy AI leaflet**.'

🔶 Co-responsible stakeholders: **Development Decisionmakers** (AI developers, data scientists, technical architects, managers etc.), considering that strategic technical decisions should be taken in collaboration between teams especially in the very early processes.

🔺 Sharing expertise/feedback: **Domain Experts (DE), AI Users (AIU) and Affectees (AFF)** (domain expertise, system users and potentially negatively affected people or groups), considering that they will be able to provide insights based on their experience or expertise. Such consultations will be valuable to include at the early stage of the AI lifecycle. They will collaborate through various participatory methods.

Sub-methodologies to be employed during this phase

🔨 Trustworthy AI Readiness Assessment (Annex I)
🔨 AI Unfairness Manifestation Database (optional)

Engagement Formats and Participatory Methods during this Phase

🗓️ 🔗 Collaboration: among **GDM & DDM** to prepare the Trustworthy AI Readiness Assessment and document the outcomes
💬 Roundtable to employ the Trustworthy AI Readiness Assessment

Outcomes of the Scoping Stage

🛏️ Draft Trustworthy AI Leaflet
🗄️ AI Unfairness Manifestations Database set-up (optional)

Process

🔗 **Prepare for the Trustworthy AI Readiness Assessment**

Who:

- Governance Decisionmakers
- Development Decisionmakers

Duration:

2-4 weeks

Considering that GDM and DDM are the co-responsible entities in this step, they will collaborate on the following points to prepare for the AI Fairness Readiness Assessment. The following steps function as a guidance for this preparation.

Identify and plan involvement of other stakeholders:

- Relevant **AFF**, **DE**, and **AIU** to the use case
- Representative in regards ethnicities, ages, gender, sexes, socio-economic classes
- Plan stakeholder involvement throughout AI lifecycle

Plan outreach to onboard AFF, DE, and AIU:

- Coordinating with or involving existing civil society partners.
- Identifying domain experts in the specific domain of application/
- Identifying (business) users of the AI system (internal and/or external)

Prepare for TAIRA:

- Identify problem holder
- Identify DDM to present AI system
- Plan and invite qualified participants to the roundtable

Set up prototype AI unfairness manifestations database (optional)

An example prototype of AI Unfairness Manifestation Database was designed as part of Deliverable 6.1 to collect cases of AI unfairness, that could inform the project, allowing categorization of information as follows:

- 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.

AFF, DE and AIU will have access at a later stage to add cases of witnessed unfairness incidents, specifying all the required details on, among others, sources, description, and type of harm resulting from the bias.

💬 Roundtable: Execution of Trustworthy AI Readiness Assessment

Who:
🔷 Governance Decisionmakers
🔶 Development Decisionmakers
🔺 Domain Experts, AI Users, Affectees

Duration:
🕛 3 hours

Sub-methodology:
🔨 Trustworthy AI Readiness Assessment (Annex I)

Outcomes:
🖨️ Trustworthy AI Leaflet

GDM will guide the roundtable starting with the **presentation of:**
- **AI lifecycle**
- **Stakeholder involvement planning**
- **Trustworthy AI Readiness Assessment (Annex I)**

GDM will then lead the execution of the Trustworthy AI Readiness Assessment (Annex I) ensuring the feedback of AFF, AIU, and DE on the AI system's fairness readiness based on the use case. Each stakeholder will be able to provide input. This is crucial as we observed in our testing of sub-methodologies that the discussion can be monopolized by a specific stakeholder group.

Risk Analysis Phase

The risk analysis stage is the second phase of the AI lifecycle where all known and foreseeable risks of unfairness should be identified, evaluated, and assessed for elimination or mitigation. The efficiency of the elimination and mitigation tools will be tested in the evaluation phase. These processes are aligned with the spirit of the Risk Management System in the EU AI Act in terms of steps and structure.

Goals during the Risk Analysis Phase:

- 📍 Co-identifying relevant **ethical**, **legal**, and **social** constraints to AI Fairness
- 📍 Establishing a Fairness Risk Management Strategy

Engaged Stakeholders during this Phase

- 🔷 Responsible stakeholders: **Governance Decisionmakers** (incl. **ethical**, **legal**, and **social** experts and managers) as they guide the process towards and within the risk analysis phase. They should make sure that all building blocks are completed, facilitate the transit from one task to another, and coordinate between the engaged stakeholders.
- 🔶 Co-responsible stakeholders: **Development Decisionmakers** (AI developers, data-scientists, technical architects, etc.), considering that strategic technical decisions that should be taken in collaboration between teams especially in assessing the technical risks and getting aware of the ethical, social, and legal constraints to consider when building the AI system.
- 🔺 Sharing expertise/feedback: **Domain Experts (DE), AI Users (AIU) and Affectees (AFF)** (domain expertise, system users and potentially negatively affected people or groups), considering that they will be able to provide valuable insights on known and foreseeable risks of the particular system taking into account the context of its deployment.

Sub-methodologies to be employed during this phase and guiding documents

- 🔨 Prohibited Social Scoring Assessment (PSSA) (Annex II)
- 🔨 Fundamental Rights Impact Assessment for Fairness (FRIA-F) (Annex III)
- 📓 Trustworthy AI Leaflet (Result of the scoping stage)
- 🛢 AI Unfairness Manifestation Database (optional)

Engagement Formats and Participatory Methods

- 🗓 🔗 Collaboration among the members of **GDM** and **DDM** to identify, evaluate, and assess the risks of unfairness and the means of addressing them.
- 📝 Contribution to Guiding Documents: **GDM** and **DDM** will co-develop the Risk Management Report where all steps of the risk analysis stage will be documented.
- 🔎 Focus Group: with **DE**, **AIU**, and **AFF** to provide comprehensive and representative insights on the risks of unfairness with a consideration of the context and potential misuses.

Outcomes of the Risk Analysis Phase

- 📓 Identification of **Ethical**, **Legal** and **Social** Constraints for AI Fairness
- 📓 Fairness Risk Management Strategy

Process

📅 🔗 Preliminary identification of Ethical, Legal and Social constraints

Who:
🔷 Governance Decisionmakers
🔶 Development Decisionmakers

Duration:
🕐 1-4 weeks

Sub-methodologies:
🔨 OECD methodology for risk level and urgency
🔨 FRIA-F (Annex III)
🔨 Prohibited Social Scoring Assessment (Annex II)

Input from Scoping phase:
📓 Trustworthy AI Leaflet

Building on the outcomes of the Trustworthy AI Readiness Assessment, documented in the Trustworthy AI Leaflet, previous GDM will start a granular analysis of the fairness risks from ethical, legal and social perspectives. This process aims to identify and assess these risks and determine a strategy to manage them.

General Risk Level and Urgency (OECD): It is important to identify the general risk level of the AI system to determine the urgency by which this risk needs to be addressed. The OECD's typology ((serious) AI incident, (serious) AI hazard, AI disaster) can be used to determine this. The OECD report on Defining AI incidents and related terms provides a clear definition of **actual** and **potential** harms.

Ethical constraints: The outcomes of ALTAI (step 4 of TAIRA), and the Ethical Lens discussion (step 5 of TAIRA) provide a concrete overview of the **ethical** considerations and constraints

Legal constraints: GDM will identify and assess the **legal** constraints for fairness considering the use case at hand. This process will be guided by the following sub-methodologies:

- 🔨 **Fundamental Rights Impact Assessment for Fairness (Annex III)** helps Governance Decisionmakers (GDM) to identify and assess the potential misalignment of the system with the fairness principles of the EU Charter of Fundamental Rights.
- 🔨 **The Prohibited Social Scoring Assessment (Annex II)** is used to identify if the system poses too great a risk to fundamental rights and would be prohibited.
- 🔨 If the system is not prohibited, a tool such as the AI Act Compliance Checker can be used to determine whether the AI system is considered high-risk under the AI Act and identify the requirements and obligations related to such classification.

Social Constraints: The outcomes of ALTAI (step 4 of TAIRA), and the Ethical Lens discussion (step 5 of TAIRA) provide a concrete overview of the **social** considerations and constraints.

GDM and DDM will closely collaborate and ensure that risks are properly labelled and assessed and agree on the risk management strategy for each one. Where risks cannot be eliminated but can only be mitigated, they will have to **agree if the residual risk is acceptable**.

A **Risk Management Strategy Report** will be drawn up comprising an identification of risks, measures for risks elimination and mitigation, description of accepted residual risks and argumentation for acceptance.

🔎 Focus Group to get feedback on the identified known and foreseeable fairness risks

Who:
- 🔷 Governance Decisionmakers
- 🔶 Development Decisionmakers
- 🔺 Affectees
- 🔺 Domain Experts
- 🔺 AI Users

Duration:
- 🕐 4 hours

Input:
- 📓 Ethical Legal and Social Constraints
- 📓 Risk Management Strategy

GDM will organize a focus group through which they will collect the informed feedback of DE, AIU, and AFF on the known and foreseeable risks of unfairness the AI system could pose, considering the specificities of the use case in hand.

GDM will present the identified fairness risks and DE, AIU and AFF will be requested to identify other risks of unfairness based on the use case and their expertise/role.

Feedback should also be collected about potential misuses and the potential need for disclaimers to support AIU & DE.

Development Phase

The development phase is the third phase of the AI lifecycle where Development Decisionmakers (DDM) are co-responsible for the completion of all processes. After identifying and analysing the risks of unfairness in the previous phase, DDM will implement the adopted risk management strategy in this phase. Choices regarding the data, the model, and the outcome interpretations are made in a shared responsibility with the Governance Decisionmakers (GDM) to make sure the identified **ethical**, **legal**, and **social** constraints are all considered.

Goals during the Development:

📍 Address and optimize AI system fairness at data, model, and outcome interpretation levels.

Engaged Stakeholders during this Phase

🔷 Responsible stakeholders: **Governance Decisionmakers** (incl. **ethical**, **legal**, and **social** experts and managers) as they guide the process towards and within the risk development phase. They should make sure that all building blocks are completed, facilitate the transit from one task to another, and coordinate between the engaged stakeholders.

🔶 Co-responsible stakeholders: **Development Decisionmakers** (AI developers, data-scientists, technical architects, etc.), considering that strategic technical decisions and execution should be in collaboration between teams. When developing an AI system, stakeholders make decisive choices on data, models, and outcomes, hence the necessity to align the system with the ethical, social, and legal constraints.

Sub-methodologies to be employed during this phase and guiding documents

🔨 Fair Data Collection, Governance, and Management (sub-)Methodology (FDCGM) (Annex IV)

🔨 Fair Model Methodology (Annex V)

🔨 Fair Outcome Interpretation Methodology (Annex VI)

🗄️ Trustworthy AI Leaflet

🗄️ Ethical, Legal and Social Constraints

🗄️ Risk Management Strategy

Engagement Formats and Participatory Methods

🔗 Collaboration among the members of **GDM** and **DDM** to implement **ethical**, **legal** and **social** constraints to fairness and follow the Risk Management Strategy

Outcomes of the Development Phase

🗄️ Data Dictionary

🗄️ Model Dictionary

🗄️ Interpretation Card/Dictionary

Process

Coordinate fairness enablers

Who:
- Governance Decisionmakers
- Development Decisionmakers

Duration:
- 2 hours

Input:
- Trustworthy AI Leaflet
- Ethical Legal and Social Constraints
- Fairness Risk Management Strategy

In this step, the goal is to help design a fair AI system critically addressing the ethical, legal, and social fairness constraints at data, model and outcome interpretation levels.

This process builds on the results of the previous phases, where ethical, legal, and social risks of unfairness were identified and analysed.

Coordinating fairness enablers is aimed at clarifying the tasks of each member of the DDM and GDM teams; DDM will be executing the technical steps enabling fair data collection, governance and management, fair model selection or design, and fair outcomes interpretation. GDM will collaborate, provide guidance and expertise and review the documented processes executed by DDM. During the coordination meeting, discussions should cover the following questions as follows:

- What are the risks of unfairness identified in the previous phases?
- Was the feedback of AIU, AFF, and DE properly included and understood?
- Where should each risk be managed: data, model, outcomes, or all?
- Which limitations of the system can already be identified (in data, model, or outcomes)?
- How will the process be documented?

The meeting will conclude with planification of the tasks of the development stage, clarification of stakeholders' roles, and agreement on the fair process to govern data, select/design the model, and interpret outcomes. The meeting details will be summarized by the GDM and/or planned in a **project management tool**.

🔗&📝 Collaborative work to ensure fairness in the data governance, model design or selection, and outcomes' interpretation

Who:
🔷 Governance Decisionmakers
🔶 Development Decisionmakers

Duration:
🕐 2-6 months

Sub-methodologies:
🔨 Fair Data Collection, Use and Governance Methodology (Annex IV)
🔨 Fair Model Selection/Design Methodology (Annex V)
🔨 Fair Interpretation Methodology (Annex VI)

Input:
📓 Fairness Risk Management Strategy

The **GDM and DDM** will co-create three documents to keep track of data collection & management, model design or selection, and model interpretation processes. They will jointly execute the **FDCGM, FMM and FOIM** and document all steps accordingly, building one document for data related processes (📘 Data Dictionary), another for model selection & design (📘Model Dictionary), and a third for outcomes' interpretation (📘 Interpretation Dictionary).

Evaluation Phase

The fourth lifecycle stage of the AI system consists of testing the collected data, the designed or selected model, and the outcomes' interpretations for fairness before deploying the system. In the evaluation stage, all stakeholders collaborate to ensure that the system is free from **ethical**, **legal**, and **social** risks.

WP6 makes the difference between **process** and **system audits** to guide stakeholders in following the steps of this stage. **Process Audits** evaluate to which extent fairness principles and values are included in the process shaping the design, development, and use of the AI system. These principles were identified, analysed, and selected in the previous stages considering the **ethical**, **legal**, and **social** constraints to fairness.

Rather than adopting a checklist to enable process audits, as observed in the state of the art[6], WP6 opts for a more comprehensive approach. Since there are diverse considerations that should be taken into account to establish AI fairness, a checklist cannot do justice to the interdisciplinary scope. Therefore, WP6 provides a **set of methodologies** to assist stakeholders in the process audit, such as the Fundamental Rights Impact Assessment for Fairness (FRIA-F) methodology for instance. These methodologies were translated into the Risk Management Report which summarizes all identified risks. All relevant documents for process audits are listed under the building blocks and tools for this stage which also specifies the methodologies for code audits. **System Audits** are the evaluation of the tools and technical components used in the development of the AI system. This evaluation represents an implementation of the selected fairness principles and covers the data, model, and outcomes of the AI system.

Goals during the Evaluation Phase

📍Assessment of the AI systems' fairness from the **ethical**, **legal**, and **social** perspectives (Process Audit):

- **Ethical**, **legal**, and **social** assessment following the Trustworthy Readiness AI Assessment methodology.
- **Legal** assessment of the AI systems based on the Fundamental Rights Impact Assessment for Fairness (FRIA-F) methodology.
- Collection of feedback from authorities.

📍Co-evaluation of the AI systems' fairness (at the data, model, and outcomes levels), the efficiency of the selected elimination/mitigation tools, and the management of potential misuses (System Audits).

Engaged Stakeholders during this Phase

🔷Responsible stakeholders: **Governance Decisionmakers** (incl. **ethical**, **legal**, and **social** experts and managers) as they guide the process towards and within the evaluation phase. They should make sure that all building blocks are completed, facilitate the transit from one task to another, and coordinate between the engaged stakeholders.

🔶Co-responsible stakeholders: **Development Decisionmakers** (AI developers, data scientists, technical architects, Tech managers, etc.), considering strategic technical decisions that should be taken in collaboration between teams, especially in evaluating the AI system and the implemented technical elimination or mitigation tools. **Supervisors-Notified Bodies** are also co-responsible in this stage as they have legal authority and/or obligation to evaluate the fairness of the AI System *(depending on the legislation in place)*.

🔺Sharing expertise/feedback: **Domain Experts (DE), AI Users (AIU) and Affectees (AFF)** (domain expertise, system users, and potentially negatively affected people or groups considering representativity in the selection), will provide valuable and comprehensive insights on known and foreseeable risks of the system taking into account the context of its deployment.

Engagement Formats and Participatory Measures

📆Meetings: among the **GDM and DDM** to legally, ethically, and socio-technically evaluate the AI system considering the fairness of data, model, outcomes' interpretation, and risk elimination or/and mitigation tools.

💬Direct contact: with Supervisors Notified Bodies to, depending on the legislation in place, inform on the technical features of the AI system, audit it for fairness, or apply for a certification before deployment.

📝Contribution to Guiding Document: the **GDM and DDM** will adapt the Risk Management Report and the three dictionaries (for data, model, and interpretation) to document the evaluation steps and keep track of modifications. The Trustworthy AI and Technical leaflets will also be updated depending on the modifications of the constraints and fairness strategy.

🔎Focus Group: with AIU, DE, and AFF to evaluate the AI system considering data, model, and outcomes. They will also provide feedback on the efficiency of the implemented tools to manage potential misuse and the efficacity of supportive tools for AIU.

Sub-methodologies to be employed during this phase and guiding documents

🔨Trustworthy Readiness AI Assessment (TAIRA)

🔨The Fundamental Rights Impact Assessment for Fairness (FRIA-F)

🔨 Fair Data Collection, Governance and Management Methodology (FDCGM)

🔨 Fair Model Methodology (FMM)

🔨 Fair Interpretation Bias Methodology (FOIM)

📓 Trustworthy AI Leaflet

📓 Ethical, Legal and Social Constraints

📓 Risk Management Strategy

📓 Data Dictionary

📓 Model Dictionary

📓 Interpretation Card/Dictionary

These tools and methodologies were already used during the Scoping, Risk, and Development Phases, and will, in this stage, be re-used to evaluate and validate the AI system's overall fairness.

Process

🔗 & 📅 Co-evaluation of AI Fairness

Who:
🔷 Governance Decisionmakers
🔶 Development Decisionmakers

Duration:
🕐 1 month

Sub-methodologies:
🔨 (Evaluate) Fair Data Collection, Use, and Management Methodology
🔨 (Evaluate) Fair Model Selection/Design Methodology
🔨 (Evaluate) Interpretation Bias Methodology
🔨 Trustworthy AI Readiness Assessment (Annex I)
🔨 FRIA-F Methodology (Annex III)
🔨 Prohibited Social Scoring Assessment (Annex II)

Input:
📓 Ethical, Legal and Social Constraints
📓 Risk Management Strategy
📓 Data Dictionary
📓 Model Dictionary
📓 Interpretation Card/Dictionary

The GDM and DDM will contribute to this step by separately and jointly assessing, testing, and evaluating the fairness of the AI system. Each member will specialize in the evaluation of the system based on her/his expertise and could work separately, but simultaneously. They will follow the tools provided and use the documentation prepared in the previous stages as a starting point.

The GDM and DDM will evaluate the effectiveness of the measures used to eliminate and mitigate the risks of unfairness. They will also evaluate the AI system based on the shared steps in the three methodologies. If any discrepancies are identified in the fairness of the data, model, and outcomes' interpretation, all changes need to be documented as a continuation of the Risk Management Report.

Revisiting the **Fundamental Rights Impact Assessment for Fairness (FRIA-F) Methodology** helps the **GDM** to identify and assess the potential misalignment of the system with the fairness principles of the EU Charter of Fundamental Rights. It will be re-executed by the **legal** experts to evaluate the AI system based on the adopted fairness principles and measures at the development stage.

The GDM will review the **three dictionaries** and the **Risk Management Report** to have an overview of the system's compliance with existing laws. This review will focus on comparing the results of the identified legal constraints in the risk analysis stage with the ones mitigated during the development phase. The review will also emphasize the fairness of followed processes at the three levels: data, model, and outcomes' interpretation. All results need to be documented in preparation for the GDM and DDM meeting.

The same process will be followed by the **social sciences** experts in evaluating the mitigation of social risks in the development of the AI system, at its three levels. The GDM will assess if the adopted measures efficiently eliminate, or mitigate, the risks of the system on language, social cohesion, religion & belief, socio-economic class, National artefacts, symbols, and value. All results need to be documented.

Revisiting the TAIRA will help the GDM assess the AI system's fairness from an ethical perspective, once developed.

The GDM will add the co-evaluation results to the **Risk Management Report**. If any changes need to be technically done, the team will make sure these modifications are established. The DDM will then update the **three dictionaries** based on the established modifications.

🔎 **RoundTable to co-evaluate the AI system with all stakeholders**

Who:
- 🔷 Governance Decisionmakers
- 🔶 Development Decisionmakers
- 🔺 Affectees
- 🔺 Domain Experts
- 🔺 AI Users

Duration:
- 🕐 3 hours

Input:
- 📓 Risk Management Report
- 📓 Data Dictionary
- 📓 Model Dictionary
- 📓 Interpretation Dictionary

🩹 **Risk Management Report** serves as a reference to all stakeholders to keep track of what has been already identified in terms of risks of unfairness and which acceptable residual risks persist.

📚 **Data, Model, and Outcomes Dictionaries** are the three created documents as a result of the development stage. These documents provide details on which risks were eliminated or mitigated in the data, model, and outcomes levels. They also give insights into which residual risks were accepted, why & how they were managed, and which supporting mechanisms were adopted to mitigate some misuse cases. These documents are the implementation of the risk management strategy and demonstrate the practical steps of fairness.

AIU, DE, and **AFF** will have access to the listed tools above to review the risk management strategy and implementation steps of the team. Then, they will provide feedback based on their experience, field of expertise, and context of the use case. The **GDM** will moderate the roundtable and make sure all stakeholders are heard. During the gathering, stakeholders will also discuss misuse cases and assess if the adopted processes are taking count of those pitfalls.

Process

The GDM will present the Risk Management Report and the **DDM** will present the AI system and the Data Dictionary, Model Dictionary and Interpretation Dictionary. The GDM will then and moderate the discussion with stakeholders, keep track of all discussed points by **DE, AIU,** and **AFF** and include them in the **Risk Management Report**. The agreed-upon modifications need to be executed by the GDM (procedural, organisational) and the DDM (technical) who will then update the **three dictionaries** accordingly.

Conclusion

This document outlined the Multi-Stakeholder Approach to AI Fairness-by-Design (MAP) that WP6 has engineered in detail to simplify the process for AI Stakeholders in co-building and maintaining a Fair AI System. We consider **known**, **foreseeable**, and **post-market** risks of unfairness which can be identified, assessed, and monitored throughout the lifecycle stages of AI systems. In this traditionally technical process, we examine **ethical**, **legal**, and **social** constraints to fairness, concretising by this approach the interdisciplinary scope of the project and highlighting its necessity in building a comprehensive FbD methodology to the existing multi-stakeholder and multi-disciplinary ecosystem.

Before diving into the roles of stakeholders and the way they will contribute to designing, developing, and deploying a Fair AI System, it is paramount to explain how they will be identified. Therefore, the Stakeholder Identification Methodology (SIM) is discussed at the start, summarizing WP6's typology of stakeholders namely, **Affectees**, **Decisionmakers**, and **Domain Experts & Users**. Each of these stakeholders is categorized into groups depending on their function or relation to the AI system. The sub-categorization of AI actors led to the identification of six groups active throughout the lifecycle stages of the system in different ways namely, **GDM**, **DDM**, **AFF**, **DE**, **AIU**, and **Supervisors-Notified Bodies**.

WP6 makes the difference between the three types of stakeholders' engagement, forming a taxonomy of their roles. **GDM**, a sub-category of decision-makers, are **responsible** for guiding the FbD process, collaborating with other stakeholders who are either **co-responsible** or **providers of domain expertise or feedback,** depending on the AI lifecycle stage. There are different steps and tasks at each stage which materialize stakeholders' contributions in several shapes. Therefore, WP6 concludes the first chapter with a taxonomy of six engagement formats highlighting the participatory methods enabling collaboration in building a Fair AI system such as focus groups and (co-)contribution to guiding documents.

After positioning this methodology in the FbD process and defining key typologies, the MAP dives directly into the process. At each AI lifecycle stage, it explains which stakeholders are engaged, their roles in the process, formats of their participation, sub-methodologies and tools to complete their tasks, documentation to track operations, and step-by-step processes to enable a Fair AI system. The provided sub-methodologies are prepared in ANNEXES I-VI and mapped in each phase to clarify and explain how they can be used in the scoping, risk analysis, development and evaluation phases of the AI system.

In conclusion, this methodology does not only explain why stakeholders should be contributing to the process but also meticulously describes how such contribution could take effect and provides the relevant methodologies and tools to guide such contribution, **merging ethical, legal, social and technical considerations of AI fairness**. This co-creative method enables the identification and management of technical, legal, ethical, and social constraints, throughout the socio-technical perspective of AI lifecycle stages namely, scoping, risk analysis, and development. It additionally allows for the inclusion of underrepresented groups in the decision-making process that is traditionally technical.