AEQUITAS
unbias AI

AI Fairness-by-Design Multi-Stakeholder Methodology

-

A Comprehensive Framework for Fair AI Design and Development

ANNEX I: Trustworthy AI Readiness Assessment Framework

“Question Zero”

Fair-by-Design Sub-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

Contents

Introduction

The **Trustworthy AI Readiness Assessment Framework** serves as the initial component of a comprehensive Multi-stakeholder Assessment Methodology for AI Fairness, developed as part of the AEQUITAS project's comprehensive approach to ensuring fairness in AI systems. The Trustworthy AI Readiness Assessment Framework addresses the critical "***question zero***" in artificial intelligence (AI) applications: determining both the necessity and readiness of AI solutions within a rigorous **ethical**, **sociotechnical** and **legal lens**. Accordingly, this assessment is conducted during the scoping phase of the AI lifecycle, so organizations can establish a strong foundation for subsequent stakeholder engagement and risk management processes.

The primary purpose of this Framework is to guide stakeholders in systematically evaluating AI systems' readiness in a way that foregrounds **ethical**, **social** and **legal** responsibility. The stakeholders involved at this stage are the **Governance Decision Makers (GDM)** and the **Development Decision Makers (DDM)**, together referred to as **Joint Decision Makers (GDM and DDM)**. It is the first step in assessing whether a system is viable and serves as a diagnostic tool to identify the **ethical**, **sociotechnical** and **legal** principles which the system must align with.

The structure is organized as follows:

1. **Problem Definition**: Establishing the scope and context of the AI application.
2. **Solution Definition**: Developing potential AI and non-AI interventions.
3. **Stakeholder Identification**: Mapping relevant stakeholders and factors influencing the problem.
4. **Ethical Assessment**: Evaluating solutions based on the EU Ethics Guidelines for Trustworthy AI.
5. **Harm Testing**: Applying sociotechnical harm assessments.
6. **Solution Selection**: Choosing the most viable solution.
7. **Reflection**: Reviewing the assessment process and its findings.

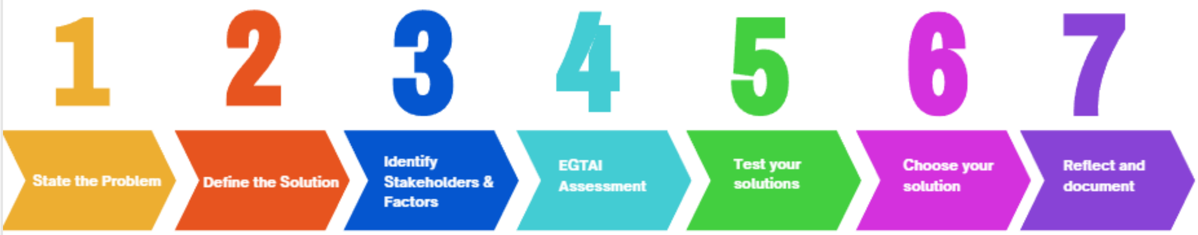


Figure 1: Seven Steps of the Trustworthy AI Readiness Assessment

Ethical evaluation within this Framework is based on the ***EU's Ethics Guidelines for Trustworthy AI***, a cornerstone document that delineates three essential pillars of Trustworthy AI: ***lawfulness*** (compliance with applicable laws and regulations), **ethicality** (adherence to principles and values), and ***robustness*** (resilience from both a technical and social perspective). These pillars serve as interdependent criteria that must be fulfilled throughout an AI system's life cycle, recognizing that each criterion is necessary, yet individually insufficient. Effective application of these principles requires them to operate synergistically; in cases of conflicting priorities, alignment efforts are essential to harmonise their impacts across society.

The **sociotechnical** perspective is embedded throughout this Framework. This perspective considers AI as part of a broader socio-institutional context where technology, social structures, and human behaviours continuously interact and influence outcomes. By highlighting these interdependencies, the sociotechnical lens acknowledges that AI systems, while designed to address specific problems, may unintentionally reinforce ***social hierarchies*** or introduce ***bias*** through abstraction errors—often when social context is overlooked or reduced to quantitative metrics. Consequently, this approach ensures that AI development accounts for the diverse and situated nature of human experience, as well as the institutional and cultural factors that shape its impact.

Bias traps within the sociotechnical perspective—such as the ***framing, portability, formalism, ripple effect***, and ***solutionism*** traps—will be integrated into the harm tests conducted in Step 5. These traps serve as critical checkpoints to identify and mitigate potential biases that may arise through abstraction, ensuring the social and technical contexts remain appropriately balanced. By actively addressing these traps, this Framework fosters a more comprehensive and fair assessment, protecting against over-simplification or de-contextualization of fairness and other social values in AI applications.

The **legal perspective** is similarly pervasive, reinforcing the Framework's foundation in regulatory standards, including the ***EU Artificial Intelligence Act***. This dual anchoring in ethical principles and legal norms ensures the Framework remains both responsive to societal values and aligned with emerging regulatory landscapes for AI governance within the European Union.

Step 1 - State the Problem

The first step in the Trustworthy AI Readiness Assessment Framework, *State the Problem*, is essential for framing AI applications within their full **ethical**, **social**, and **legal** contexts. This phase defines the purpose and scope of the AI solution by carefully analysing the problem it aims to address.

Objectives

The primary objective of this phase is to precisely delineate the problem, focusing on three core aspects:

1. **Defining the problem scope** to ensure AI is appropriate as a solution.
2. **Establishing context** to account for social, cultural, and institutional factors.
3. **Identifying prohibited systems** to avoid investing unnecessary resources into systems that should not be developed

Defining the problem scope

Scoping the problem is crucial to avoiding a purely technological approach to issues that may have non-technical elements requiring nuanced solutions. This component involves:

- Delineating the specific aspect of the problem that may be addressed with AI, outlining what falls within and beyond the system's purview.
- Assessing whether AI is the **optimal tool** for addressing this problem, or if alternative or complementary solutions might provide a more balanced response.

Questions that may be considered in this phase include:

- What is the primary problem the AI system is intended to address?
- What specific tasks or challenges within this problem can be effectively addressed by AI technology?
- Why is AI being considered as the solution for this problem? What unique value does it bring compared to other approaches?
- Are there elements of this problem that may be better addressed by non-technical or policy interventions?
- What are the clear boundaries of this AI application (e.g., where does its role start and end)?

Establishing context

AI systems do not operate in isolation but are deeply embedded within the **socio-political, cultural**, and **institutional** settings where they are deployed. A thorough contextual analysis is thus critical for understanding the broader environment influencing the problem. This entails examining:

- The societal and organizational needs that drive this AI application.
- Historical, cultural, and economic factors shaping the issue.
- Existing social inequalities or structural challenges that could impact the AI solution's effectiveness and fairness.

Analysing these elements is essential to fully integrate the socio-technical lens into the Framework. Situating the problem within its full social context allows stakeholders to evaluate not only whether an AI system could effectively address the problem but also understand how suitable it is for the broader environment within which it will function.

Questions that can prompt reflection on contextual elements are:

- What are the historical, economic, or cultural factors that shape this problem?
- How have societal values or norms surrounding this issue evolved, and how might these influence the reception of an AI solution?
- Are there any notable cultural dynamics (e.g., trust in technology, skepticism about data usage) that might affect the solution's acceptance or impact?
- Are there existing social inequalities (e.g., related to race, gender, socioeconomic status) that might influence how different groups are impacted by this AI system?
- How might the AI solution risk reinforcing existing disparities, and what precautions could mitigate this?
- Additional factors (aimed specifically at the organisation) include: political/internal pressure, finances, available budget, available personnel, etc timeline.

Identifying Prohibited Systems

Before proceeding with detailed problem definition and scoping, it is essential to first determine whether the proposed AI system falls under any of the prohibited categories established by Art. 5 of the EU AI Act. This initial screening helps organizations avoid investing resources into solutions that are fundamentally prohibited under EU law.

Key questions to ask include:

1. **Manipulation and deception**

- Does the system use subliminal techniques beyond conscious perception?
- Does the system employ manipulative or deceptive techniques that materially distort behaviour?
- Does the system impair individual ability to make informed decisions?

2. **Exploitation of vulnerabilities**

- Does the system exploit age-related vulnerabilities?
- Does the system exploit disability-related vulnerabilities?

- Does the system exploit social or economic situation vulnerabilities?

3. Social scoring systems

- Does the system evaluate or classify persons based on social behaviour?
- Does the system use personal or personality characteristics gathered in another context than the current context?
- Can the use of such personal or personality characteristics lead to detrimental treatment in the current context?

4. Predictive policing

- Does the system make risk assessments to predict likelihood of criminal offenses?
- Is the system based solely on profiling or personality trait assessment?
- Exception: Does the system support human assessment based on objective facts?

5. Facial recognition

- Does the system create or expand facial recognition databases?
- Does the system use untargeted scraping of facial images?
- Does the system collect from internet or CCTV footage?

6. Emotion recognition

- Does the system employ emotion recognition in workplace settings?
- Does the system employ emotion recognition in educational settings?

7. Biometric categorization for sensitive characteristics

- Does the system infer race, ethnicity, political opinion, trade union membership, sexual orientation, religious beliefs?

8. Real-time, remote biometric identification for law enforcement

- Does the system use real-time remote biometric identification in public spaces used for law enforcement purposes?

If the answer to any of the above questions is positive, the system is likely to be prohibited as per Art. 5. The assessment should conclude at this phase, as further evaluation would be unnecessary. Only after confirming that the system is not prohibited should the assessment proceed to detailed problem scoping and context analysis.

Step 2 - Define the Solution

After thoroughly defining the problem and its context in Step 1, Define the Solution serves to develop a range of possible interventions that address the identified challenges. The primary objective of this phase is to facilitate structured brainstorming among GDM and DDM regarding potential solutions, which will be assessed against **ethical, social** and **legal** considerations in later steps.

The phase mainly requires GDM and DDM to:

- ***Generate a list of solutions*** that consider AI, non-AI, and hybrid approaches.
- ***Analyse practical considerations*** detailing how the proposed solution would solve the problem
- ***Conduct a socio-technical assessment*** around the systems practical implementation.

Generate a list of solutions

This step encourages GDM and DDM to design a range of solutions to ensure that the final decision is based on a variety of perspectives. By proposing AI-based, non-AI, and hybrid solutions, GDM and DDM can identify the unique strengths and limitations of each approach.

- ***AI-based Solutions*** may focus on automation, advanced data analysis, or predictive capabilities, which can offer efficiency and precision but may require sophisticated infrastructure and oversight.
- ***Non-AI Solutions*** could include policy changes, human-led processes, or other strategies that achieve similar results without the need for complex technology, often valuable when the solution requires a more human-cantered approach.
- ***Hybrid Solutions*** combine AI and non-AI elements, leveraging AI where it provides added value while maintaining human oversight or input where necessary.

Analyse Practical Considerations

The second element of solution definition involves evaluating how the proposed solution would solve the identified problem and highlighting practical considerations around its implementation.

- ***Problem-Solution Alignment Evaluation***: begins by explaining how the solution directly addresses the problem defined in Step 1, including a clear outline of the proposed system's function, intended outcomes, and rationale for its design.
- ***Practical Implementation Considerations:*** This includes evaluating the technical infrastructure needed, such as data access and storage requirements, computational resources, and scalability. Additionally, considerations around training requirements, user engagement, and resources for continuous improvement are critical to gauge the feasibility of implementation.

Socio-Technical Matrix Assessment Framework

Following the initial generation of potential solutions and preliminary analysis of practical considerations, it becomes essential to conduct a comprehensive **socio-technical** assessment of each AI-based solution. This assessment serves to surface and examine the underlying assumptions, epistemological foundations, and implicit claims that underpin proposed technological interventions. Notably, if the **epistemological claims** upon which the system is built are biased, unfair or based on pseudo-science, the system cannot be evaluated based on whether it functions ‘as intended’. Thus, a **socio-technical** matrix assessment framework is proposed, wherein the systematic evaluation process ensures that solutions are constructed upon sound theoretical and practical foundations before advancing to stakeholder analysis phases.

Taking inspiration from Chowdhury et al.'s (2022)[1] work on socio-technical assessment frameworks for algorithmic systems, we have developed a simplified matrix to help GDM and DDM document and reflect on the core aspects of their proposed AI solution. This streamlined template is designed to capture essential information that can inform later phases of ethical assessment and technical development. The template is shown in Figure 2; being a template not all columns are filled in. The socio-technical matrix comprises six fundamental analytical dimensions:

System Identification and Contextual Positioning: The assessment begins with a delineation of the proposed AI system's scope and operational context. This includes articulating its primary functional role within the broader organizational ecosystem and establishing its position within relevant implementation phases. This positioning must account for both technical capabilities and organizational integration patterns.

Goal Architecture and Success Parameters: This dimension articulates the system's objectives, that align with organizational requirements and strategic objectives.

Data Infrastructure and Knowledge Requirements: At the scoping phase of the AI lifecycle, this dimension simply requires the GDM and DDM o identify where the data for their AI system would be collected from. More extensive evaluations of data in **ethical, social** and **legal** contexts will occur at later phases of the AI lifecycle.

Functional Architecture and Processing Paradigms: This component provides a clear, non-technical description of how the system processes information to generate its outputs or recommendations.

Assumption and Theoretical Underpinnings: Here the GDM and DDM re required to identify and examine the fundamental assumptions underlying the system's operational logic and decision-making processes. Providing empirical evidence for and theoretical justification of these assumptions is encouraged.

Epistemological Foundation Analysis: This dimension requires an examination of the theoretical and philosophical foundations underlying the proposed solution. The analysis might include:

- Historical precedent evaluation
- Scientific validation assessment
- Theoretical critique examination
- Contemporary relevance analysis

[1] Sloane, M., Moss, E., & Chowdhury, R. (2022). A Silicon Valley love triangle: Hiring algorithms, pseudo-science, and the quest for auditability. *Patterns*, *3*(2), 100425. https://doi.org/10.1016/j.patter.2021.100425

Element	Questions	Information
System Identification and Positioning	What specific tasks and functions will your system perform?	
	How will it integrate with existing organizational workflows?	
Goals	What concrete outcomes should the system achieve?	
	How will you measure success?	
Data	What data will your system need?	
	Where will this data be collected from?	
Function	In plain language, how does your system convert input data into outputs?	
Assumption	What key assumptions does your system make about relationships between inputs and outputs?	
	What evidence supports these assumptions?	
Epistemological Roots	Where do the assumptions made by your system come from?	
	What is their intellectual lineage?	
	Are there any critiques or limitations of these theoretical foundations?	

Figure 2: Socio-technical Matrix

Step 3 - Identify Relevant Stakeholders and Factors

Building on the solutions developed in Step 2, Step 3 identifies all stakeholders who may be affected by or involved in the proposed AI system. Engaging a diverse set of stakeholders is crucial to ensuring the AI solution aligns with the needs, values, and concerns of all impacted parties.

Key objectives of this step are to:

- **Stakeholder Identification** mapping out the key stakeholder groups that should be considered.
- **Stakeholder Engagement** determining the appropriate level and method of engagement for each stakeholder group.

Stakeholder Identification

The first task is to comprehensively identify if indeed all relevant stakeholder groups, which can broadly be situated within the following categories, are identified and engaged:

- **Affectees (AFF):** Individuals or groups who will be directly or indirectly impacted by the AI system, whether positively or negatively. This includes end-users as well as marginalized or vulnerable communities.

- **Decision Makers (DM):** Individuals or organizations responsible for the development, deployment, and governance of the AI system. This encompasses AI developers, project managers, organizational leadership, and regulatory bodies.
- **Domain Experts (DE):** Specialists who can provide expert insights on the technical, social, ethical, and legal considerations relevant to the AI system. This may include ethicists, legal scholars, social scientists, and industry experts.

To identify the relevant stakeholders, the table proposes that the following questions can be asked:

AI Stakeholder Group	Identifying Questions	List of stakeholders
Affectees (AFF) (Stakeholders affected by AI systems)	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?	
Decision Makers (DM) (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? 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?	
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?	

Figure 3: Stakeholder Identification

Stakeholder Engagement Plan

Having identified the key stakeholders, it is crucial to map out the roles that the various stakeholders play throughout the life cycle of the AI systems. This ensures effective collaboration across all phases of the AI Lifecycle to create fair-by-design systems.

Interdisciplinary teams must work together in clearly defined roles to effectively implement fairness principles and mitigate bias. This structured approach not only promotes accountability and transparency but also helps organisations avoid potential legal liabilities from AI-related harm.

AI Lifecyle

The AI Lifecycle can be divided into six phases:

Phase 1: Scoping:

The scoping phase establishes the fundamental parameters of the AI project, beginning with a clear definition of objectives and an evaluation of their alignment with EU regulations. A thorough assessment determines whether an AI solution is both necessary and proportional to the identified needs, while a fairness readiness evaluation examines organizational capabilities to develop and maintain fair AI systems. The phase includes identifying key stakeholders who will be affected by or involved in the system's development and deployment and establishes the ethical, legal, and social parameters that will guide the project's development.

Phase 2: Risk Identification:

Building on the scoping phase findings, the risk analysis phase systematically identifies potential risks and biases that could affect the AI system, examining historical biases, potential data biases, and algorithmic biases that could emerge during development and deployment. This phase thoroughly evaluates known and foreseeable risks to fairness, considering ethical implications, legal compliance requirements, and potential social impacts across different stakeholder groups. The analysis culminates in developing specific risk management strategies that outline prevention measures, mitigation techniques, and monitoring protocols for addressing identified risks throughout the system's lifecycle.

Phase 3: Development

The development phase encompasses data generation, modelling, and interpretation phases, representing the core technical implementation and human interaction aspects of the AI system:

A. **Data Generation**

Data Collection

The process begins with identifying the target population of people or phenomena for study. Features and labels must be carefully defined and measured, though it's rarely feasible to include an entire population. Instead, practitioners typically work with a development sample, a representative subset of the target population. Many AI practitioners opt to use existing datasets rather than undertake new data collection, which introduces its own set of fairness considerations and potential biases.

Data Preparation/Curation

Once collected, data undergoes various pre-processing steps to ensure quality and usability. The dataset is typically split into training data for model development and test data for evaluation. A portion of the training data is usually reserved for validation purposes. This phase requires careful attention to maintain representativeness and avoid introducing or amplifying biases through preprocessing decisions.

B. **Modelling**

Model Selection/Development

During this phase, models are selected or built using the prepared training data. The process focuses on optimizing specified technical objectives, such as minimizing mean squared error, through various model types and optimization techniques. The choice of model architecture and parameters must balance technical performance with fairness considerations.

Post-processing

After training, models undergo various refinement steps to enhance their performance and fairness. This includes implementing bias mitigation techniques and adjusting model outputs as needed to meet both performance and fairness criteria. Post-processing provides a crucial opportunity to address any biases that emerge during training.

C. **Interpretation**

Human-AI Interaction

The interpretation phase focuses on how humans interact with and understand the AI system's outputs. Testing must account for automation bias, where users over-rely on automated systems, and confirmation bias, where users favour information that confirms their preexisting beliefs. Processes must be developed to address undesirable outcomes that may arise from these interactions.

Transparency and Explainability

Interpretability mechanisms and clear explanation methods are essential for ensuring users can understand system decisions. This includes comprehensive documentation of system behaviour and limitations. Transparency measures help users make informed decisions about when and how to rely on system outputs.

Human Bias Prevention

Special attention must be paid to identifying and preventing human biases in result interpretation. This involves developing safeguards against biased interpretations and establishing clear guidelines for proper result interpretation. A robust process for challenging questionable outputs ensures ongoing oversight and improvement of the system's fairness.

Phase 4: Evaluation

The evaluation phase serves as a critical pre-deployment checkpoint where the system's performance is comprehensively assessed against established fairness criteria. This may involve several strategies, such as testing the system's behaviour across different demographic groups, validating bias mitigation strategies, and ensuring compliance with ethical, legal, and social requirements defined in earlier phases.

Phase 5: Monitoring

During deployment, the AI system transitions from controlled testing environments to real-world applications, requiring vigilant monitoring of its performance and impact. This phase implements continuous surveillance mechanisms to detect emerging biases, unfair outcomes, or unexpected system behaviours that may not have been apparent during testing. Active monitoring may involve strategies such as tracking fairness metrics, collecting stakeholder feedback, and maintaining oversight of system decisions to ensure they align with intended fairness goals and do not amplify existing societal inequities.

Phase 6: Re-evaluation

The re-evaluation phase implements an iterative approach to system assessment, regularly revisiting the comprehensive evaluation procedures established in Step 5. This ongoing process examines how the system's performance and fairness metrics may have evolved , incorporating new insights from deployment experiences and stakeholder feedback.

Please note that phases 5 and 6 will not be covered, as the MAP focusses on the ‘design’ phases of the AI system only.

Stakeholder engagement in AI Lifecycle

Within the AI lifecycle, stakeholders can be engaged in the following ways:

- ***Responsible*** (leading or guiding the process)
  - Lead and guide processes
  - Make final decisions
  - Ensure accountability
  - Coordinate between stakeholder groups
- ***Co-responsible*** (actively assisting responsible stakeholder)
  - Actively assist responsible stakeholders
  - Provide technical or domain expertise
  - Implement decisions and recommendations
  - Support coordination efforts
- ***Provider of Domain Expertise or Feedback***
  - Offer consultative input
  - Provide specialized knowledge
  - Share experience-based insights
  - Represent affected communities

The level of involvement of each stakeholder may change throughout the AI lifecycle, thus it’s important to clearly map what kind of engagement is needed and expected at each phase. Thus, the final phase of Step 3 Identify Relevant Stakeholders is to have GDM and DDM complete a stakeholder engagement map as seen in the table below.

AI stakeholders	Scoping	Risk	Development	Evaluation	Deployment Monitoring	Re-evaluate

Affectees **(AFF)**	🔺	🔺		🔺	🔶	🔺
Governance Decision Makers **(GDM)**	🔷	🔷	🔷	🔷	🔷	🔷
Development Decision Makers **(DDM)**	🔶	🔶	🔶	🔶	🔶	🔶
Supervisors Notified Bodies				🔶	🔶	
Domain Experts **(DE)** & Fairness Feedback Providers **(FF)**	🔺	🔺	🔶/ 🔺	🔺	🔶	🔺
AI Users **(AIU)**	🔺	🔺	🔺	🔺	🔶	🔺

🔷Responsible
🔶Co-responsible
🔺Sharing expertise/feedback

Figure 4: Stakeholder Engagement in Lifecycle

Relevant Factors Identification

When developing AI systems, various factors can influence both the definition of the problem and the development of potential solutions. These factors can act as either supporting elements that facilitate implementation and success, or as inhibiting elements that create barriers and challenges. The following section provides an overview of potential factors that may affect the broader context in which an AI system will operate. This list is not exhaustive, it is imperative to consider other contextual factors that may affect the problem's scope, definition and emerging solutions.

Organisational Resources

Organisational resources are the material and non-material factors that affect how a problem is conceived and addressed. These may include things such as the allocation of funding; current (technical) infrastructures that are in place; the availability of staff; staff skillsets; and the work culture.

Questions to consider:

- How do departmental budgets influence which aspects of the problem are prioritised?
- How does the organizational culture affect which problems are seen as "worth solving"?
- In what ways might staff expertise and background influence how the problem is understood?
- What technical infrastructure and systems are currently available?

- How do current staffing levels and expertise align with potential solution requirements?
- What financial resources could be allocated for development and maintenance?

Political Context

Political factors encompass both external political systems that shape how problems are defined, and which AI systems are developed and deployed. These might include elements such as the political salience of issues; which problems are prioritised and funded in current policies; citizens' opinions on the use of AI in specific contexts; and geopolitical pressures.

Questions to consider:

- How do current political priorities influence which aspects of the problem receive attention?
- What role does public opinion play in how the problem is understood and framed?
- How do geopolitical tensions or relationships affect which problems are considered urgent?
- What is the current political climate regarding AI deployment in this domain?
- How might upcoming policy changes or elections affect project support and funding?
- What level of public trust exists for AI systems in this context?

Legal Framework

Legal considerations establish the boundaries within which AI systems must operate. Understanding the legal context is crucial for ensuring compliance and managing potential risks throughout the system's lifecycle.

Questions to consider:

- What national laws and regulations apply to the proposed system?
- How do current data protection and privacy laws impact development?
- What industry-specific regulations must be considered?
- What are the legal requirements for transparency and accountability?

Step 4 - Assess AI solution(s) against the Ethics Guidelines for Trustworthy AI (EGTAI)

Step 4 of the methodology involves conducting a systematic assessment of the proposed AI solution against the Ethics Guidelines for Trustworthy AI (EGTAI). This critical evaluation phase requires stakeholders to examine whether their AI solution meets the proposed ethical requirements. Additionally, as the ethical considerations are situated within the broader socio-technical context and thus the sociological perspective is integrated through-out the evaluation.

GDM and DDM should consider all requirements thoughtfully, even when certain aspects appear less relevant to their specific AI solution. Initial impressions of relevance may overlook subtle but important connections that emerge through deeper analysis. Requirements that seem peripheral at first often reveal significant considerations during a thorough examination. The assessment process should be used as an active tool for improvement rather than a compliance checklist. Any issues or concerns identified during this evaluation should prompt GDM and DDM to revise and refine their proposed solutions.

The following section provides an overview of all seven of the EGTAI requirements whilst situating them in the broader **sociotechnical** lens. Guiding questions are also provided to aid GDM and DDM in the assessment of their systems. To better streamline the assessment process however, we recommend the use of the Assessment List for Trustworthy AI (ALTAI) **Platform**. ALTAI was designed by the High-Level Expert Group on Artificial Intelligence set up by the European Commission to help assess whether an AI system meets the EGTAI requirements. The platform takes users through a series of questions that help them identify which **ethical** elements need further attention. The final output of the assessment is visualised in a spider graph as seen in Figure 5. ALTAI also provides users with recommendations on how to best address the systems shortcomings as exemplified by the image in Figure 5.

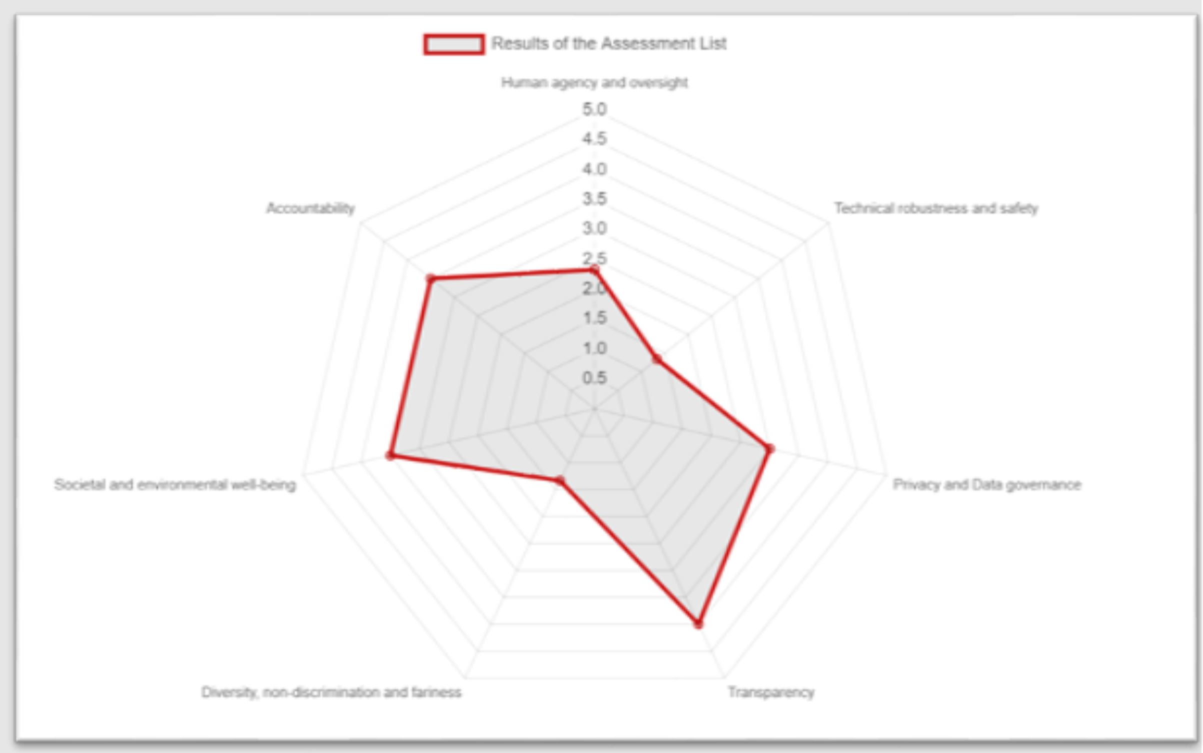
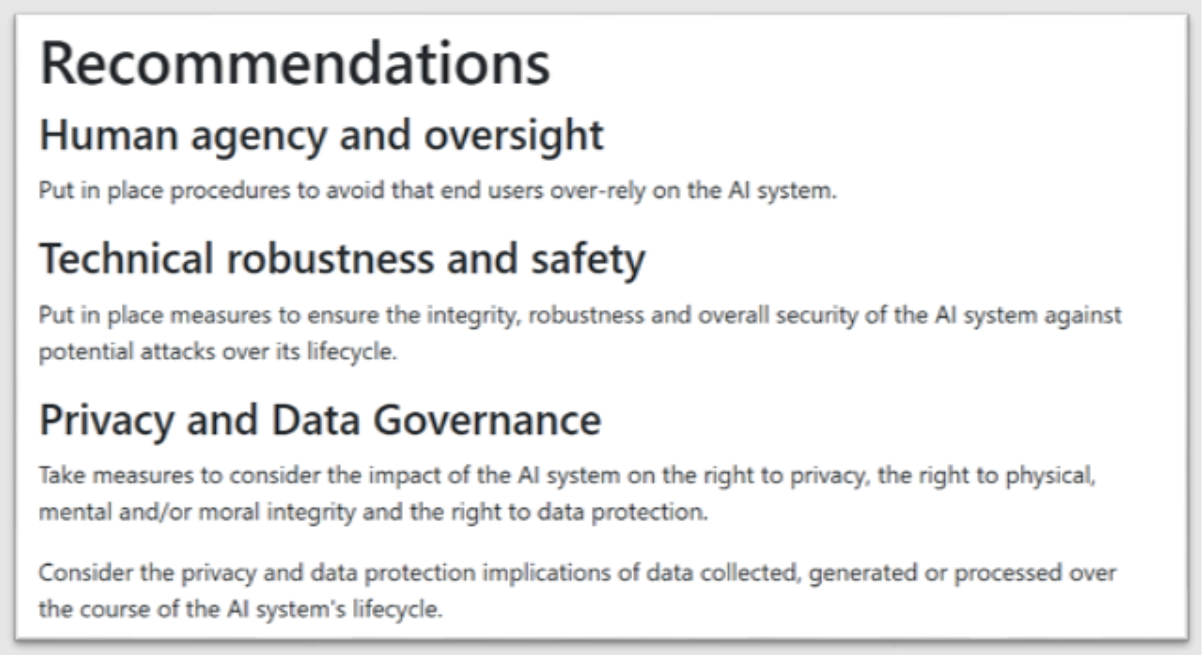


Figure 5: Sample EGTAI Spider graph



Recommendations

Human agency and oversight

Put in place procedures to avoid that end users over-rely on the AI system.

Technical robustness and safety

Put in place measures to ensure the integrity, robustness and overall security of the AI system against potential attacks over its lifecycle.

Privacy and Data Governance

Take measures to consider the impact of the AI system on the right to privacy, the right to physical, mental and/or moral integrity and the right to data protection.

Consider the privacy and data protection implications of data collected, generated or processed over the course of the AI system's lifecycle.

Figure 6: Sample EGTAI Recommendations

Additionally, the following section highlights several **sociotechnical** traps that AI systems are prone to and which must be addressed, namely: ***the portability trap, the framing trap, the formalism trap and the ripple effect trap***. Accordingly, it is important to examine the ethical implications of these systems within the context of these traps, thus guiding questions have been provided to help GDM and DDM evaluate their solutions.

Key requirement #1: Human Agency and Oversight

Human Agency and Oversight addresses the fundamental role of AI systems as enablers of informed human decision-making within a democratic and equitable society. Such systems must support individual agency, protect fundamental rights, and incorporate appropriate human oversight mechanisms.

Three distinct oversight approaches may be implemented:

- **Human-in-the-Loop (HITL):** Enables direct human intervention within individual decision cycles. While offering maximum control, this approach may not be feasible or beneficial in all contexts.
- **Human-on-the-Loop (HOTL):** Facilitates human intervention during system development and enables ongoing operational monitoring, providing oversight without requiring constant direct intervention.
- **Human-in-Command (HIC):** Grants comprehensive oversight of system activities, including broader economic, societal, legal, and ethical impacts. This approach empowers decision-makers to determine appropriate system deployment, including the authority to withhold system use in specific situations.

Each approach offers different levels of human agency and control, appropriate for different contexts and applications.

The following questions could serve as a guidance:

Human agency:

- Does the AI system enhance or augment human capabilities?
- Is this AI system human-centric: does it leave meaningful opportunities for human choice?
- Does it enable individuals to have more control over their lives or does it limit their freedom and autonomy?

Human oversight:

- Can you describe the level of human control or involvement in your solution?
- Have you considered a ‘human-in-the-loop’ or a ‘human-on-the-loop’ or a ‘human-in-command’?
- Does the human performing the oversight have the relevant skills, knowledge and authority?

Key requirement #2: Technical Robustness and Safety

Technical Robustness and Safety emphasises the development of AI systems with proactive risk management strategies. Such systems must demonstrate reliable and reproducible performance, actively minimising unintentional and unexpected harm, and preventing unacceptable harm. The system design must account for potential environmental variations and interactions with other entities, both human and artificial, including possible adversarial behaviours. Additionally, attention must be given to protecting both physical and psychological human wellbeing.

The following questions can serve as a guidance:

- Resilience to attack and security: Can you identify any potential forms of attacks to which the AI system could be vulnerable?
- Fallback plan and general safety: Is there a probable chance that the AI-system may cause damage or harm to users or third parties?
- Accuracy: How could the accuracy of the system be measured and ensured?
- Reflect: If your solution does not meet this requirement, is there a way to adapt it accordingly?

Beyond the technical robustness analysis, in order to ensure the safety of AI systems, it’s essential to consider the social and institutional context from within which they operate. To do so effectively, it’s important to consider and address the bias traps identified within the **sociotechnical** perspective (Selbst et al. 2019). Within the Technical Robustness and Safety analysis step, the **portability trap** is most relevant to address.

Portability Trap

The **sociotechnical** lens posits that concepts such as ‘fairness’ or ‘trustworthiness’ are not inherent to the system itself, but rather to the specific social contexts from which it operates. Accordingly, while data or a system may be fair within a specific context, when applied in a different setting may introduce bias and unfairness.

The portability trap in algorithmic systems can be clearly demonstrated through the case of a lending algorithm originally developed for Country A's context, where a robust social safety net, high financial literacy, and standardized credit reporting inform its notion of "fairness." In this original setting, the algorithm's script interprets missed payments as genuine financial distress, employment gaps as voluntary breaks, and bank account fluctuations as reliable indicators of financial health—all reasonable assumptions within Country A's highly formalized economic structure.

However, when this same algorithm, carrying its embedded "fair" label, is transported to Country B, its script fundamentally misaligns with the local context, where informal economies thrive, many citizens operate outside the traditional banking system, and financial arrangements often follow different social norms. For instance, what the algorithm interprets as negative indicators (such as missed formal payments or employment gaps) may actually reflect perfectly sustainable informal economic arrangements in Country B's context.

This transplantation illustrates how the attempt to make algorithmic fairness portable across contexts fundamentally misunderstands that fairness is not an intrinsic property of code but rather a context-dependent assessment that must be evaluated against specific social, economic, and cultural backgrounds—a nuance that is lost when developers fall into the trap of believing that a "fair" algorithm will retain its fairness across different societal contexts.

Questions to consider the **portability trap** may include:

- When we say our algorithm is 'fair' which specific cultural and social assumptions are we unconsciously embedding from our context?
- How might our understanding of 'normal user behaviour change if we moved this technology from an urban to a rural setting? From one country to another?
- What local practices or cultural norms could make our current script assumptions inappropriate or even harmful in a different context?
- What social, economic, or cultural factors in different contexts might require us to fundamentally rethink our fairness/ethic/trustworthiness metrics?

Key requirement #3: Data Privacy and Governance

Requirement 3 - Privacy and Data Governance recognises privacy as a fundamental right that requires protection in AI systems, given their inherent dependence on data. As these systems frequently process information about individuals, inadequate data stewardship can have significant personal impacts. Privacy protection must therefore be prioritised through comprehensive data governance frameworks.

Thus, the proposed solutions must be evaluated against multiple dimensions: data quality and integrity, contextual relevance for the specific domain of deployment, robustness of access protocols, and privacy-preserving processing capabilities.

To do so effectively, the following questions may be asked:

- Respect for privacy and data protection: Are there ways to develop the AI-system or train the model without or with minimal use of potentially sensitive or personal data?
- Quality and integrity of data: Can you think of oversight mechanisms for data collection, storage, processing and use?
- Access to data: What protocols, processes and procedures can you think of to manage and ensure proper data governance

Key requirement #4: Transparency

Transparency addresses multiple facets of system openness and understandability. It encompasses the distinction between transparent and opaque systems, the ability to explain system decisions and reasoning, and clarity regarding all relevant elements of the AI system. This includes transparency about data collection, system training, operational processes, and business models. Comprehensive documentation and communication of design decisions and data usage are essential to achieve this requirement.

Whilst evaluating transparency aspects of their proposed solutions, GDM and DDM can use the following guiding questions:

Traceability:

- What mechanisms could you establish that facilitate the system's auditability, such as ensuring traceability and logging of its processes and outcomes?
- Did you review the outcomes of or decisions taken by the system, as well as potential other decisions that would result from different cases (for example, for other subgroups of users)?

Explainability:

- Can you explain why the system will make a certain choice in a way that is understandable for all users?

Communication:

- What mechanisms can you put in place to inform (end-)users on the reasons and criteria behind the AI-system's outcomes
- What is the exact purpose of your AI-system and who or what may benefit from it?
- Can you specify usage scenarios for the system and clearly communicate them to ensure that the system is understandable and appropriate for the intended audience?

Key requirement #5: Diversity, non-discrimination and fairness

Diversity, Non-discrimination and Fairness acknowledges that AI systems, despite their foundation in data and logic, are not inherently objective or unbiased. The data that underpins these systems reflects human choices in collection methods, measurement parameters, and selection criteria, inevitably incorporating human subjectivity. Achieving trustworthy AI therefore requires active consideration of diversity and inclusion throughout the system's lifecycle. This encompasses both comprehensive stakeholder engagement and the implementation of inclusive design principles to ensure equitable access and treatment for all users.

The following considerations can serve as a guidance:

Unfair bias avoidance:

- Assess and acknowledge the possible limitations stemming from the composition of the used data sets.
- Assess whether there could be persons or groups who might be disproportionately affected by negative implications.
- Accessibility and universal design.
- Assess whether the AI system is usable by those with special needs or disabilities or those at risk of exclusion. How can this be designed into the system and how can it be verified?

Stakeholder participation:

- Can you think of ideas to include the participation of different stakeholders in the AI system's development and use?

When considering diversity, non-discrimination and fairness, the **framing and formalism traps** from the **sociotechnical** perspective should also be considered. Assessing whether an AI system falls into these traps will allow GDM and DDM to further understand how inherently subjective human choices can introduce bias or unfairness into technological solutions.

Framing Trap

The framing trap occurs when complex **sociotechnical** problems are incorrectly defined as purely technical challenges, leading to incomplete solutions that overlook crucial social, institutional, and human factors. It's the error of drawing system boundaries too narrowly around technical components while excluding the social context in which they operate. This overly narrow framing typically results in solutions that may be technically sound but fail to address the full scope of the problem or effectively in real-world conditions where social and technical elements are inevitably intertwined. This can lead to discrimination and unfair outcomes.

Imagine a hospital re-admission prediction system to illustrate how narrow technical framing can limit the effectiveness of AI solutions in healthcare. In its most basic implementation, engineers develop a machine learning model that predicts which patients are likely to be re-admitted within 30 days of discharge. This model typically focuses solely on medical data, including diagnoses, vital signs, lab results, and previous hospital visits, ultimately generating a readmission risk score from 0-100.

However, this narrow technical framing overlooks crucial social factors that significantly influence readmission rates. These include the patient's ability to afford follow-up care, access to reliable transportation for appointments, availability of family members to assist with recovery, and potential language barriers with medical staff. Additionally, institutional factors such as hospital staff workload, resource constraints, local pharmacy access, and insurance coverage limitations play critical roles in determining patient outcomes.

A more effective approach adopts a broader sociotechnical framing that incorporates both technical and social elements. This expanded framework includes not only the ML model but also considers hospital protocols, staff training, and community resources. It examines how the prediction system interacts with existing hospital workflows, accounts for varying resource levels across hospital departments, factors in local community support systems, and incorporates feedback mechanisms from medical staff and patients.

Questions to identify the **framing trap** might include:

- What elements are we currently including in our system definition?
- What social and human factors are we treating as "external" to the system?
- Where have we drawn the line between what's "in" and "out" of our solution space?
- Are we artificially separating technical and social components that actually work together?

Formalism Trap

The Formalism Trap occurs when people assume that because something can be defined mathematically or formally, that definition fully captures and solves the underlying social problem. It's particularly dangerous because once a mathematical definition is accepted, people may stop questioning whether the system produces fair or just outcomes in real-world contexts.

Consider a city's automated school assignment system. The developers create an algorithm that defines "fairness" mathematically as: every student gets assigned to one of their top 3 school choices while maintaining racial and economic diversity at each school.

The technical team celebrates this as an elegant solution, and the city adopts it, considering the fairness problem "solved" because they have a mathematical model ensuring "fair" distribution. However, this formalism traps them into missing crucial real-world factors:

1. Some families can't easily transport their children to distant schools
2. The "top 3 choices" assume all families have equal access to information about schools
3. School quality differences aren't addressed
4. Community ties and neighbourhood cohesion aren't factored in

While the system achieves its mathematical definition of fairness, it may actually create new inequities for low-income families who need schools close to home or work. The formalism (mathematical model) has masked the deeper social justice issues rather than solving them.

The trap isn't that the mathematical model is wrong - it's that accepting it as a complete solution prevents people from seeing and addressing the broader social context and real-world impacts.

Questions to address the **formalism trap** may include:

- What aspects of fairness are we unable to quantify in our mathematical model?
- What assumptions are embedded in our mathematical definitions?
- Whose definition of fairness are we encoding into these formulas?
- Whose voices or perspectives might be missing from our formalization process?
- How would different communities or groups define this problem differently?

Key requirement #6: Societal and Environmental Wellbeing

Societal and Environmental Well-being positions both environmental and societal welfare as essential stakeholders throughout an AI system's lifecycle. This requirement extends beyond immediate system impacts to consider broader effects on democratic processes and public discourse, while promoting sustainability and ecological responsibility. It encompasses both the potential for AI to address societal challenges such as climate change, and the responsibility to minimise the ecological impact of AI system development and deployment. Implementation requires both interdisciplinary expertise and comprehensive impact assessment methodologies.

Whilst evaluating societal and environmental aspects of their proposed solutions, GDM and DDM can use the following guiding questions:

Societal impact:

Ensure that the social impacts of the AI system are well understood, by asking for example the following questions:

- Could the AI system negatively influence or polarise public discourse?
- Could the AI system negatively affect democratic processes or democracy?
- Could the AI system negatively affect the judiciary, and the rule of law?

- Could the AI system be used to manipulate or confuse people?
- Assess whether the AI system could contribute to substantial job loss or skills erosion.

Sustainable and environmentally friendly AI:

Ensure that the environmental impact of the system is well understood, by asking for example the following questions:

- What mechanisms could you establish to measure the environmental impact of the AI-system's development, deployment and use? (For example, think about the type of energy used by the data centres).
- What measures can you think of that can reduce the environmental impact of your AI-system's life cycle?

From the **sociotechnical** perspective, it is essential to address the **ripple effect trap** in this step to further contextualise the effect the AI system will have on society at large.

The Ripple Effect Trap

The ripple effect trap occurs when introducing new technology appears to simply modernise or optimise existing organizational processes but triggers a cascade of unintended social and political consequences that reshape power dynamics. While the technology might seem neutral on the surface, it can become a catalyst that either reinforces existing power structures or creates new tensions between different organizational groups. This trap is particularly insidious because the ripple effects often manifest gradually and can be difficult to anticipate during the initial implementation phase.

Imagine the implementation of AI-powered scheduling at a General Hospital. Initially introduced as a neutral optimization tool to reduce administrative burden and ensure fair shift distribution, the system quickly became entangled in existing power dynamics and sparked unexpected social consequences.

While senior physicians and administrators championed the system for its supposed objectivity, its design quietly reinforced existing hierarchies by granting override privileges to senior staff and weighting their preferences more heavily in the algorithm. The system's implementation triggered a cascade of reactive behaviours: nurses discovered the algorithmic bias toward physician preferences and began coordinating their availability inputs to game the system; departments manipulated patient flow data to justify higher staffing levels; and staff learned to exaggerate their constraints to achieve desired outcomes.

As these dynamics unfolded, different stakeholder groups reinterpreted the technology's purpose through their own lenses – administrators came to view it as a cost control mechanism, senior physicians as a tool for preserving their privileges, and nursing staff as an instrument of hierarchy enforcement. The HR department's traditional role in mediating scheduling conflicts diminished, while the IT department gained unexpected authority as system administrators.

This transformation exemplifies how a seemingly neutral technological intervention can be "torqued" from its original purpose of promoting fairness to instead reify existing power structures and create new forms of organizational tension. The case demonstrates how technological implementation must be understood not just as a technical challenge but as an intervention into complex social and political dynamics that can produce far-reaching and often unexpected ripple effects throughout an organization's social fabric

When considering the **ripple effect trap**, **GDM and DDM** may reflect on the following questions:

- Who currently holds decision-making power in the affected domains?
- Which groups might gain or lose influence through this technology?
- How might access controls and permission levels reinforce or challenge existing hierarchies?
- How might the technology affect democratic processes and civic participation?
- What new forms of social capital might emerge or be diminished?
- How might this technology reshape relationships between institutions and citizens?
- How might this technology alter social norms and cultural practices? What new cultural factors may emerge?

Key requirement #7: Accountability

Accountability mandates the establishment of clear responsibility and accountability mechanisms for AI systems throughout their entire lifecycle - from development through deployment and ongoing use. This encompasses comprehensive auditing processes, thorough record-keeping practices, and appropriate legal liability frameworks. System developers and operators must demonstrate their commitment to and success in minimising potential negative impacts.

When assessing accountability measures within their proposed solutions, GDM and DDM should address the following:

Auditability:

- What mechanisms could you establish that facilitate the system's auditability, such as ensuring traceability and logging of its processes and outcomes?

Minimising and reporting negative impact:

- Carry out a risk or impact assessment of your AI system, taking into account different stakeholders that are (in)directly affected.
- Think of processes that you can establish for third parties (e.g. suppliers, consumers, distributors) or workers to report potential vulnerabilities, risks or biases in the AI-system.

Documenting trade-offs:

- What are the relevant interests and values impacted by the AI-system?
- What are the potential trade-offs between them?
- How do you decide on such trade-offs? Document the trade-off decision.

Ability to seek redress:

- What mechanisms can you establish to allow for redress in case of the occurrence of any harm or adverse impact?

Step 5 - Test Your AI Solutions

Step 5 introduces a series of ethical assessment tests designed to evaluate the broader societal implications and impacts of AI solutions. These tests provide a structured framework to examine potential consequences, public perception, defensibility, and professional integrity of proposed systems. By systematically addressing societal, ethical, and professional dimensions, these tests help ensure that AI solutions not only meet technical requirements but also align with broader social values and responsibilities.

Whilst evaluating their solutions, GDM and DDM should address the following tests:

Harm test - Does the solution do less harm than the other solutions?

- Does one solution take some criteria better into account than others?
- Is it possible to combine the best of different solutions in one solution?
- Is your solution necessary to solve the problem and is it limited to solving the problem?
- Does your solution pay specific attention to vulnerable groups and ensure that they are not treated with bias?
- Would you support this solution if you were the one adversely affected and harmed?

Publicity test - Would I want my solution published in the newspaper?

- What sort of questions and concerns from the public would your solution raise?
- Does the solution take into account all the relevant stakeholders? How does it advantage or disadvantage some stakeholders over others?
- Does the solution have some broader societal impact?
- How democratic is the solution? Consider the effect on agency and the power of citizens.

Defensibility test - Could you defend choosing this solution to a governmental committee, a committee of my peers or my parents?

- Is the solution **lawful** (legally allowed)? In the EU context, it is important to consider whether the solution is **prohibited** under the EU AI Act.
- Does the solution undermine human rights (e.g. life, safety, privacy, non-discrimination, freedom of information, freedom of demonstration, a healthy and safe workplace, fair trial, etc.)?
- What is the reasoning behind choosing this solution over others and can I defend that reasoning?

Virtue test - How does your solution reflect you?

- What kind of beliefs, assumptions, attitudes, and values does your solution reflect?
- What kind of beliefs, assumptions, attitudes, and values does the process of selecting your solution reflect?
- What kind of values and ideals do you want to promote with your solution?
- Was the solution chosen independently or does it serve someone's interests?

Professional test - What might an ethics committee say about your solution?

- Does it promote the ethics of the field?
- What would your peers, classmates, colleagues say about the ethical alignment of your solution?
- What would your superior say when you describe the problem and suggest this solution?

Step 6 - Choose your Solution

Solution Selection requires GDM and DDM to synthesize their analyses from Steps 4 and 5 to determine the most appropriate and trustworthy solution from their proposed options. This critical decision-making phase evaluates both AI and non-AI solutions against established criteria for trustworthiness, effectiveness, and ethical impact.

To enable GDM and DDM to better evaluate each solution, they should fill in the following tables in Parts A through D of this section, clearly noting down how each solution meets or falls short of the requirements evaluated in the previous steps of the assessment, allows for adequate comparison and solution selection. Additionally, whilst not present here in table form, the results of the ALTAI assessment should be considered and analysed in this step.

Part A: Solution Overview

Criteria	Solution 1	Solution 2	Solution 3
Brief Description			
Type (AI/Non-AI/Hybrid)			
Implementation Complexity (Low/Medium/High)			
Resource Requirements			
Primary Stakeholders			

Part B: ALTAI Spider Graph

The ALTAI spider graph and recommendations can be inserted here.

Part C: Traps Overview

Criteria	Solution 1	Solution 2	Solution 3
Is my solution susceptible to the Portability Trap?	**(Yes – Partial – No)**	**(Yes – Partial – No)**	**(Yes – Partial – No)**
Is my solution susceptible to the Framing Trap?	**(Yes – Partial – No)**	**(Yes – Partial – No)**	**(Yes – Partial – No)**
Is my solution susceptible to the Formulism Trap?	**(Yes – Partial – No)**	**(Yes – Partial – No)**	**(Yes – Partial – No)**
Is my solution susceptible to the Ripple Effect Trap?	**(Yes – Partial – No)**	**(Yes – Partial – No)**	**(Yes – Partial – No)**

Part D: Ethical Lens Overview

Criteria	Solution 1	Solution 2	Solution 3
Harm Test			
Overall Harm Reduction	**(Yes – Partial – No)**	**(Yes – Partial – No)**	**(Yes – Partial – No)**

Impact on Vulnerable Groups	(Yes – Partial – No)	(Yes – Partial – No)	(Yes – Partial – No)
Necessity & Proportionality	(Yes – Partial – No)	(Yes – Partial – No)	(Yes – Partial – No)
Publicity Test			
Public Perception	(Positive – Mixed – Negative)	(Positive – Mixed – Negative)	(Positive – Mixed – Negative)
Stakeholder Balance	(Yes – Partial – No)	(Yes – Partial – No)	(Yes – Partial – No)
Democratic Impact	(Yes – Partial – No)	(Yes – Partial – No)	(Yes – Partial – No)
Defensibility Test			
Prohibited	(Yes – Partial – No)	(Yes – Partial – No)	(Yes – Partial – No)
Legal Compliance	(Yes – Partial – No)	(Yes – Partial – No)	(Yes – Partial – No)
Human Rights Impact	(Positive – Mixed – Negative)	(Positive – Mixed – Negative)	(Positive – Mixed – Negative)
Virtue Test			
Value Alignment	(Yes – Partial – No)	(Yes – Partial – No)	(Yes – Partial – No)
Ethical Promotion	(Yes – Partial – No)	(Yes – Partial – No)	(Yes – Partial – No)
Professional Test			
Field Ethics Alignment	(Yes – Partial – No)	(Yes – Partial – No)	(Yes – Partial – No)
Peer Review Likelihood	(Yes – Partial – No)	(Yes – Partial – No)	(Yes – Partial – No)
Professional Standards	(Yes – Partial – No)	(Yes – Partial – No)	(Yes – Partial – No)

Part E: Qualitative Analysis

Criteria	Solution 1	Solution 2	Solution 3
Key Strengths			
Primary Concerns			
Risk Level (Low/Medium/High)			
Implementation Challenges			
Long-term Sustainability			

Step 7 - Reflect & Document

Step 7 of the Trustworthy AI Readiness Assessment Framework comprises two essential components:

1. Systematic reflection on the evaluation process and outcomes of the Trustworthy AI assessment
2. Documentation of findings through the AEQUITAS Trustworthy AI Leaflet

Reflection Process

The reflection component constitutes a critical element of the assessment framework, serving as more than a mere retrospective review. This systematic examination of the assessment process enables GDM and DDM to consolidate their learning, challenge their assumptions, and identify potential oversights or biases in their analysis.

By critically examining how the understanding of the problem and potential solutions evolved throughout the assessment, GDM and DDM can identify where initial assumptions may have limited their perspective and where new insights emerged. This reflection helps surface any unconscious biases, reveals the effectiveness of the assessment methodology itself, and builds institutional knowledge for future AI system evaluations.

The reflection component requires GDM and DDM to critically examine their journey through the assessment process by considering the following dimensions:

Problem Understanding

- Has the perception of the initial problem evolved through the assessment process?
- Does the problem appear different following detailed analysis?
- Were the original assumptions accurate and well-founded?

Solution Assessment

- What alternative solutions emerged during the evaluation process?
- Were non-AI or non-technological solutions adequately considered?
- Could different approaches have been explored more thoroughly?

Organizational Context

- What preventive measures might avoid similar challenges in future implementations?
- How can institutional support be strengthened to facilitate implementation?
- What structural or organizational changes might be beneficial for system success?
- What additional resources or expertise would enhance implementation capabilities?

Documentation Process

The culmination of the Trustworthy AI Readiness Assessment methodology requires documentation of all analytical elements to be filled in the Trustworthy AI Leaflet. This documentation synthesizes the findings from the preceding steps and establishes a formal record of the assessment process.

The Trustworthy AI Leaflet will serve as a foundational document within the broader Fair-by-Design (FbD) methodology of the AEQUITAS project. It functions as a critical reference point for subsequent phases of the AI lifecycle and helps ensure a systematic approach to implementing fairness principles across all phases of AI system development.

Using the evaluation conducted in **Step 6 Choose your Solution**, GDM and DDM will be asked to fill out the following information in the Trustworthy AI Leaflet as shown in the Annex.

System Information

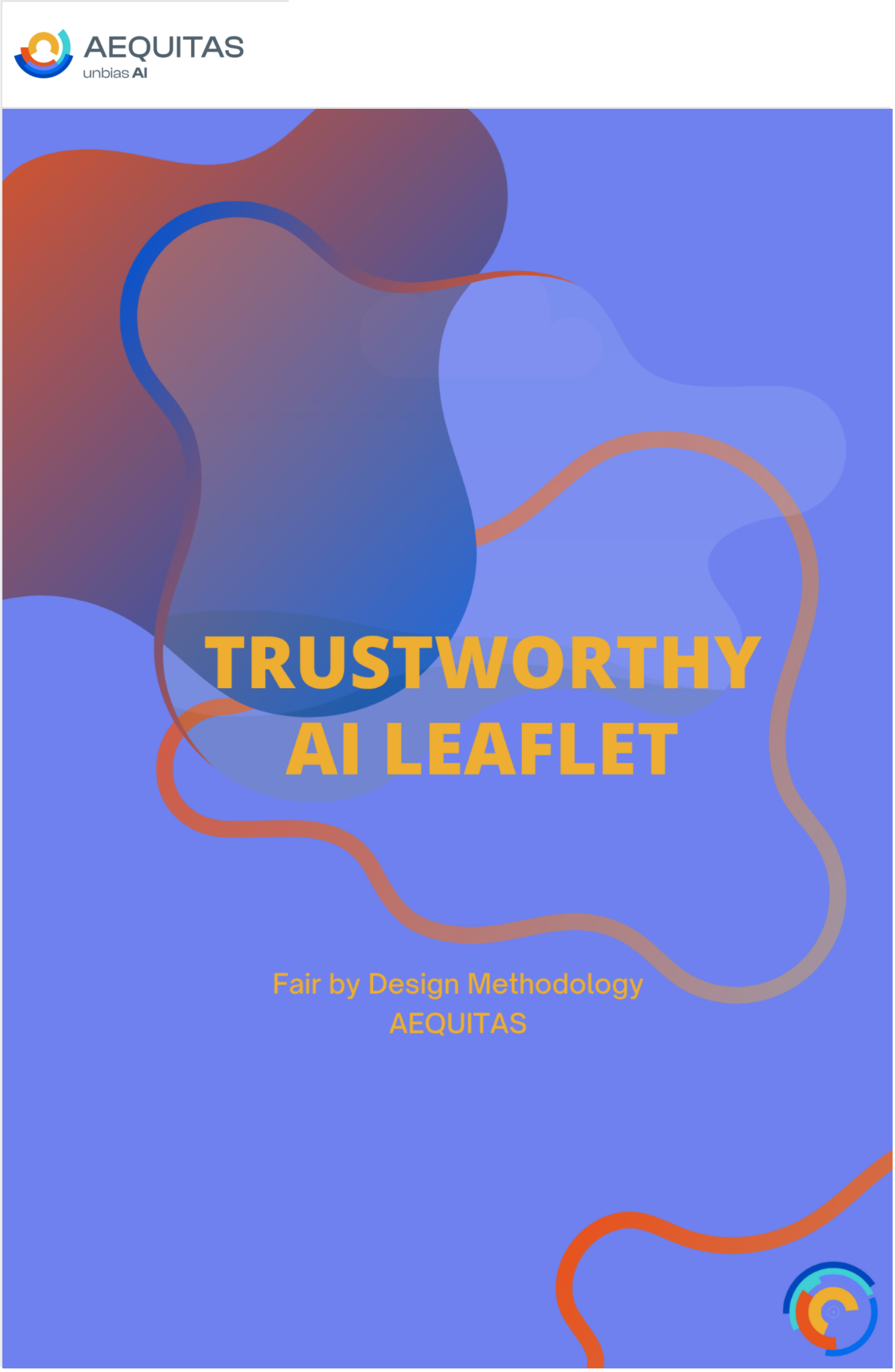
- Name and detailed description of the AI system
- Technical classification and system type
- Intended purpose and objectives
- Risk classification according to the AI Act
- Key strengths
- Key concerns
- Implementation Challenges
- Long-term sustainability

ALTAI Assessment & Risk Management Steps

In **Step 4: Assess your AI solution(s) against the Ethics Guidelines for Trustworthy AI (EGTAI)** of the process, GDM and DDM will have conducted an ALTAI assessment of the proposed AI solution. The resulting spider graph and ALTAI recommendations should be transferred to the leaflet, documenting specific measures for each requirement.

Results of Ethical Assessment Tests

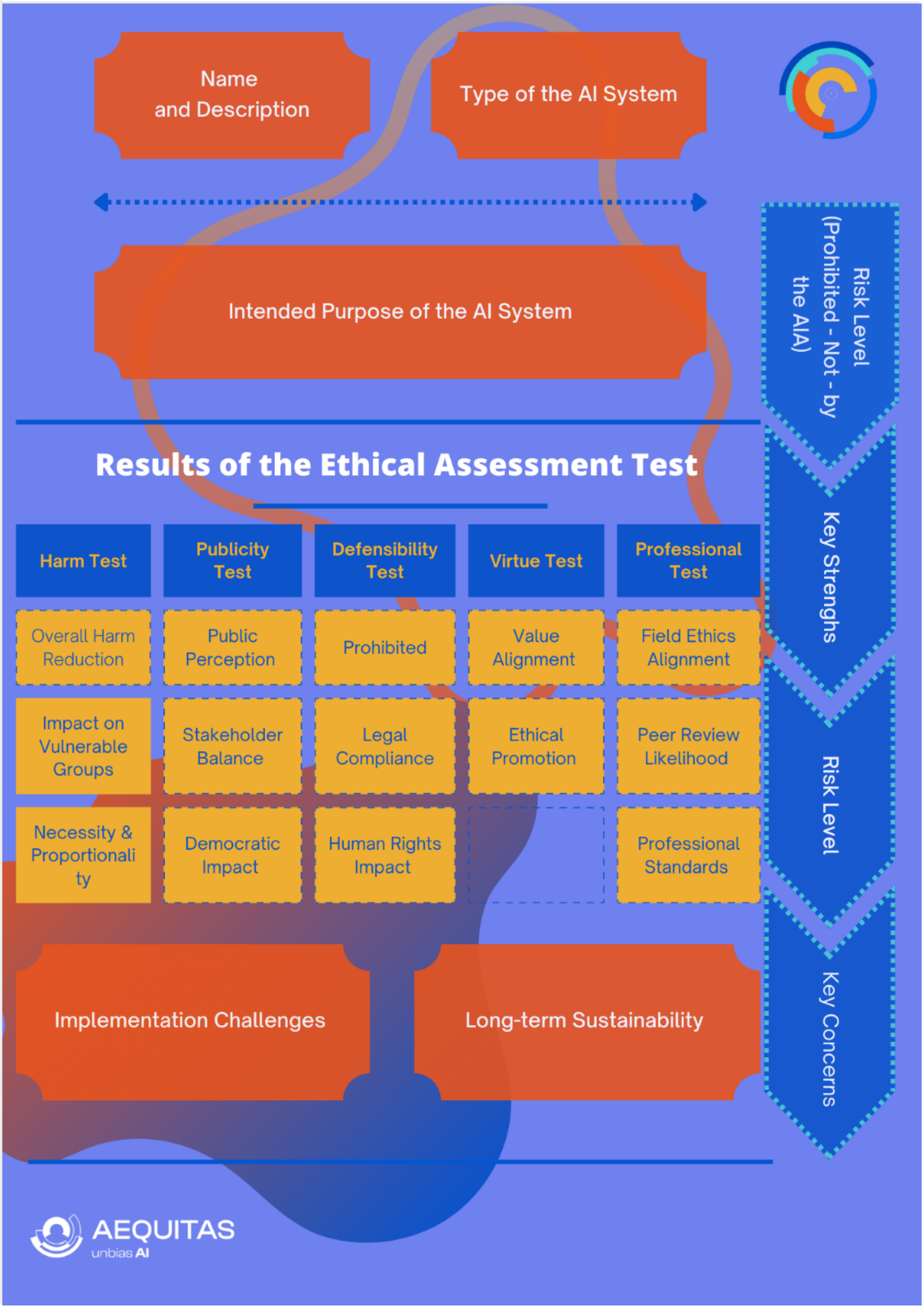
1 Harm Test: Overall harm reduction, impact on vulnerable groups, necessity and proportionality
2 Publicity Test: Public perception, stakeholder balance, democratic impact
3 Defensibility Test: Legal compliance, human rights impact, reasoning justification
4 Virtue Test: Value alignment, ethical promotion
5 Professional Test: Field ethics alignment, professional standards, peer review assessment



AEQUITAS
unbias AI

TRUSTWORTHY AI LEAFLET

Fair by Design Methodology
AEQUITAS



Name and Description

Type of the AI System

Intended Purpose of the AI System

Results of the Ethical Assessment Test

Harm Test	Publicity Test	Defensibility Test	Virtue Test	Professional Test
Overall Harm Reduction	Public Perception	Prohibited	Value Alignment	Field Ethics Alignment
Impact on Vulnerable Groups	Stakeholder Balance	Legal Compliance	Ethical Promotion	Peer Review Likelihood
Necessity & Proportionali ty	Democratic Impact	Human Rights Impact		Professional Standards

Implementation Challenges

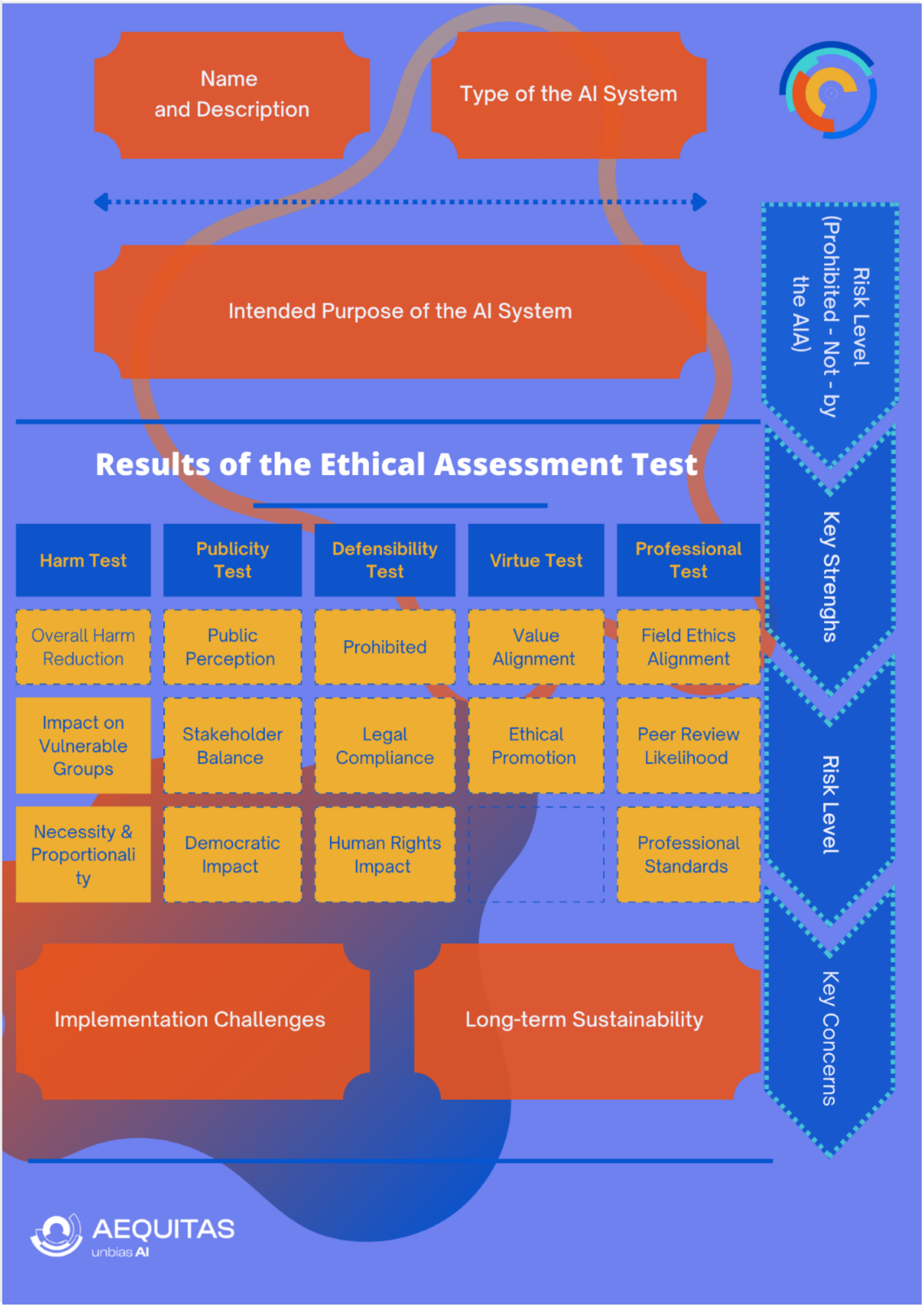
Long-term Sustainability

Risk Level (Prohibited - Not - by the AIA)

Key Strenghs

Risk Level

Key Concerns



Name
and Description

Type of the AI System

Intended Purpose of the AI System

Results of the Ethical Assessment Test

Harm Test	Publicity Test	Defensibility Test	Virtue Test	Professional Test
Overall Harm Reduction	Public Perception	Prohibited	Value Alignment	Field Ethics Alignment
Impact on Vulnerable Groups	Stakeholder Balance	Legal Compliance	Ethical Promotion	Peer Review Likelihood
Necessity & Proportionality	Democratic Impact	Human Rights Impact		Professional Standards

Implementation Challenges

Long-term Sustainability

Risk Level (Prohibited - Not - by the AIA)

Key Strenghs

Risk Level

Key Concerns