Deliverable 5.5
Report and SoS Addressing Organizations

07-04-2025
Version 1.0

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Agency. Neither the European Union nor the granting authority can be held responsible for them.

PROPERTIES

DISSEMINATION LEVEL	PUB - Public
VERSION	1.0
STATUS	Final Version
BENEFICIARY	WLC
LICENSE	 This work is licensed under a Creative Commons Attribution-No Derivatives 4.0 International License (CC BY-ND 4.0). See: https://creativecommons.org/licenses/by-nd/4.0/

AUTHORS		
	Name	**Organisation**
Document leaders	Magdalena Góralczyk	WLC
	Silvina Pezzetta	WLC
	Lucrezia Nicosia	WLC
Participants	Georgia Panagiotidou	AUTH
Reviewers	Tanja Zdolsek Draksler	JSI
	Celia Parralejo Cano	DPB

VERSION HISTORY				
Version	**Date**	**Author**	**Organisation**	**Description**
0.1	03-03-2024	Magdalena Góralczyk, Silvina Pezzetta, Lucrezia Nicosia	WLC	Initial draft version and ToC
0.2	04-09-2024	Magdalena Góralczyk, Silvina Pezzetta, Lucrezia Nicosia	WLC	Draft version for discussion
0.3	07-10-2024	Magdalena Góralczyk, Silvina Pezzetta, Lucrezia Nicosia	WLC	Second draft version for discussion
0.4	10-11-2024	Magdalena Góralczyk, Silvina Pezzetta, Lucrezia Nicosia	WLC	Annex included in the Deliverable
0.5	27-02-2025	Magdalena Góralczyk, Silvina Pezzetta, Lucrezia Nicosia	WLC	Internal WLC reviewed completed
0.6	25-03-2025	Celia Parralejo Cano	DPB	1st Internal Review
0.7	26-03-2025	Tanja Zdolsek Draksler	JSI	2nd Internal Review
0.9	29-03-2025	Silvina Pezzetta	WLC	Final version for submission
1.0	07-04-2025	Spiros Borotis	MAG	Final editing and submission

Table of Contents

Table of Figures

Abstract

Deliverable 5.5 is providing tools for public, private, or third party institutions that finance projects to develop systems, services, or products that include AI. This Deliverable is part of WP5, which aims to create and disseminate materials to raise public awareness of the impact of AI on the consolidation of democratic values. In particular, this Deliverable is focused on a cut of the large audience, the funding bodies, following the grant agreement. Moreover, the Deliverable includes ethical aspects, checklists, parameters, and criteria that funding bodies should use to evaluate applicants' proposals. Deliverable 5.5 includes two self-assessment tools. One is the "Stop and Think" tool for applicants to implement the necessary steps to comply with ethical and legal requirements. The second one, addressed to funding bodies, is the "Statement of Support," which allows application evaluators to decide on projects' ethical and legal strengths and allocate resources accordingly. In order to meet the objectives of this deliverable, the document is divided into three chapters. The first one is devoted to the ethical role of funding bodies. The second chapter introduces criteria, parameters, and a checklist for funding bodies. The last chapter provides the two self-assessment tools to help funding bodies and applicants develop trustworthy AI systems.

1 Introduction

1.1 Purpose and scope

The concerns about the potential harms of AI, especially those related to discrimination and bias, and how to prevent or redress them through ethical and legal regulation is part of what is considered a problematic tension to resolve. It is the problem of innovation versus regulation. The aforementioned concerns about the harms that the development and use of AI may cause have generated a robust body of ethical principles from experts, companies, and government agencies (Kluge Corrêa, et al., 2023). In turn, this concern gained legal status with the passage of the AI Act.

This tension between regulation and innovation centres on two main actors: states and private companies. However, little attention has been paid to the role of funding bodies for AI system development projects in achieving a positive balance between the two extremes. Thus, D5.5 – as written in the grant agreement – focuses on the funding processes for AI research grants, which we have identified as a gap in the landscape of ethical AI solutions, such as AI procurement guidelines, AI impact assessments, and AI audit frameworks. D5.5 emphasizes the responsibility of funding bodies to direct investments toward the development of trustworthy and safe AI systems.

The main focus of this Deliverable is the "Trustworthy AI Statement" section within grant applications. D5.5 assumes a "Trustworthy AI Statement" is part of every call for applications, and it is the key first step to ensure ethically and legally aligned AI projects. Therefore, this Deliverable will provide a set of criteria, checklist, and parameters to evaluate the "Trustworthy AI Statement" and decide which are worth funding. Two self-assessment tools are part of D5.5 to ensure the achievement of its goals. These tools are a "Stop-and-Think" document for applicants to use when preparing their applications and their "Trustworthy AI Statement" and a "Statement-of-Support" (SoS) document for funding bodies to evaluate applications.

1.2 Document structure

The Deliverable is structured into three chapters. The first one presents the legal and ethical framework and key discussions surrounding AI challenges to democracy and human rights, with a special focus on the role of funding bodies. This chapter will briefly review the most important ethical guidelines and the AI Act. The second chapter introduces ethical parameters, a checklist, and funding bodies' criteria for assessing applications. The third chapter presents the self-assessment tools, the "Statement-of-Support" for funding bodies, and the "Stop-and-Think" tool for applicants.

1.3 Target audience

The primary intended target audience of the Deliverable are the key funding bodies across the consortium countries. Funding bodies include government agencies, private sector companies, and non-profit organizations. This Deliverable also includes a self-assessment tool for funding applicants. This means the audience includes individuals and organizations seeking funding to develop AI projects. However, other possible interested stakeholders include:

Project Stakeholders: Other stakeholders engaged in the AI4Gov project, including external advisors, experts, and policymakers, may find this deliverable valuable, as *it provides checklists, parameters, criteria, and self-assessment tools that can be adapted for various purposes.*

Researchers and Academia: This deliverable may be relevant to researchers and academics specializing in AI, governance, ethics, and fundamental rights.

Policy and Decision-Makers: They may find D5.5 relevant, as both the report and self-assessment can be helpful to them.

2 The Role of Funding Bodies in Ensuring Trustworthy AI

2.1 Introduction

Though AI has evolved over time, its recent surge into fields like healthcare and education raises concerns about the implications and potential impacts. Moreover, the sudden and widespread adoption of AI systems, especially generative models like ChatGPT, has led to significant debate over their impacts on human rights, the environment, democracy, and human jobs, to name the most important topics (Coeckelbergh, 2024). The way to address the potential harms includes seeking compliance of AI systems with ethical principles and, more recently, in the EU, with legal regulation through the AI Act.

A key concept in all these approaches, particularly in the EU, is "trustworthy AI." Trustworthy AI, thus, is the backbone of ethical guidelines and the AI Act. Trustworthy AI encompasses a detailed and systematic framework crucial for individuals and societies. This approach ensures that ethical standards, transparency, and accountability are prioritized when developing, deploying, and utilizing AI systems. By fostering an environment of trust, we can harness the full potential of AI technologies while safeguarding human rights and preventing harm to humans and the environment (Díaz-Rodríguez et al., 2023).

As Díaz-Rodriguez et al. explain (2023), the trustworthy AI approach is made of three pillars: ethical, legal and technical (Figure 1).

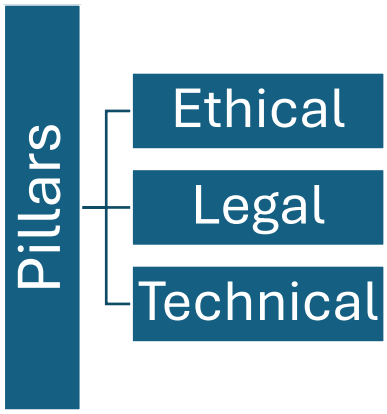

Figure 1: AI Pillars (Díaz-Rodriguez et al., 2023)

These pillars are the foundation of seven requirements to develop a trustworthy AI system: the legal, ethical, and technical robustness pillars; and the following requirements: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination, and fairness; societal and environmental well-being; and accountability[1].

[1] European Commission High-Level Group on AI, Guidelines for Trustworthy AI, 2019

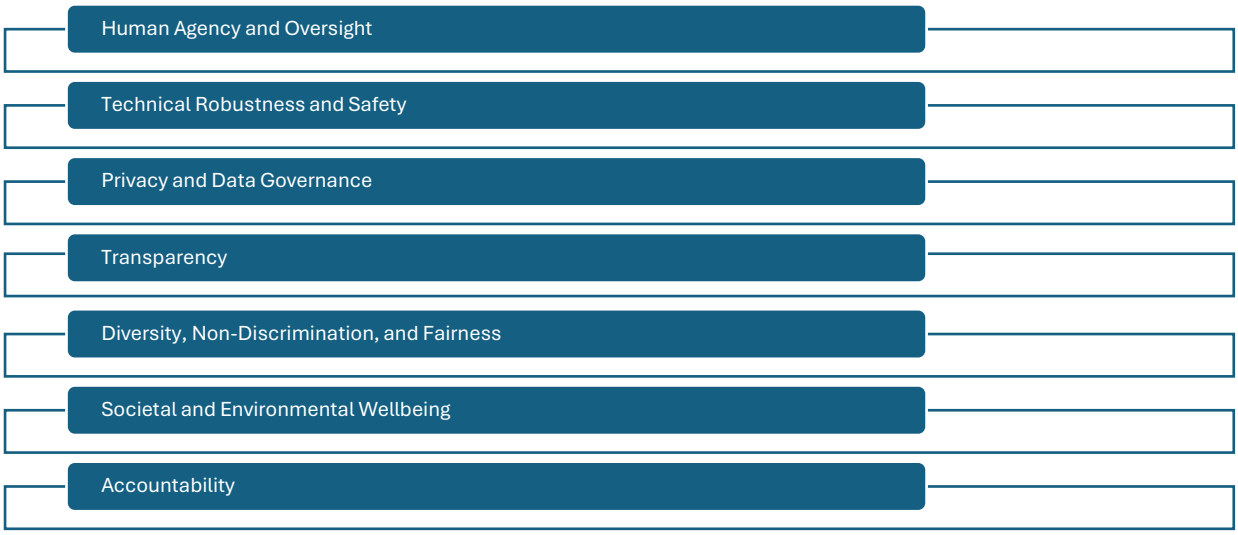

Figure 2: Requirements

In turn, "Trustworthy AI" became one of the main foundations of the AI Act. However, the EU is not the only region that has proposed ethical guidelines and regulations. A brief account of key guidelines and approaches will be presented in the following sub-section. Finally, in the last sub-section of this chapter, D5.5 offers an analysis of the key elements of the AI Act.

2.2 Ethics Guidelines

2.2.1 ALTAI

During the 2010s, the US and China led the race to develop and harness AI. In this scenario, the EU sought to position itself to avoid being left behind. The first result of the response to the race for AI was the AI Strategy of April 2018[2]. The strategy sought to differentiate itself from China and its focus on the development of state-oriented AI, as well as the pro-market orientation of the US (Smuha, 2024).

As explained in deliverable D1.4 (delivered in M12), the EU's position was to become a leader based on its values. Thus, the EU Commission proposed that "we can make a difference in the approach to AI that benefits people and society as a whole." To this end, it set up a group of experts, the High-Level Experts Group (HLEG)[3], through a public call to create a guide of ethical principles for developing and using AI. Thus, the first approach of the EU to the AI challenges was through the development of soft law and not regulation (hard law). In addition, not all experts in the EU were worried about the risks and potential damages of AI systems. However, as will be noted soon in this deliverable, during the ethics guidelines discussions and with the passing of the AI Act, concerns about harm would be one of the main focuses.

[2] https://ec.europa.eu/commission/presscorner/detail/en/ip_18_3362
[3] https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

The term "Trustworthy AI" is a probe that shows that the conversation about AI has turned into one about its potential perils. As Natalia Smuha (2024) says:

"The High-Level Experts Group (HLEG) opened the Guidelines with a description of what it meant with the term 'Trustworthy AI,' of which the advancement and promotion is considered the document's main aim. While the Commission requested the preparation of 'ethics guidelines for AI', it did not suggest the use of the term 'Trustworthy'. This term was deliberately chosen by the HLEG after internal deliberations on what the most suitable concept would be, and subsequently also added to the document's title".

It is worth quoting the way in which the HLEG refers to the centrality of "Trustworthy AI":

"we … identify Trustworthy AI as our foundational ambition, since human beings and communities will only be able to have confidence in the technology's development and its applications when a clear and comprehensive framework for achieving its trustworthiness is in place"[4].

The aforementioned guidelines provide a structured approach to achieve ethical and robust Trustworthy AI. Chapter I of the HLEG guidelines establishes the foundational ethical principles, drawing from various philosophies like Kantian, utilitarian, and virtue ethics, ultimately settling on a fundamental rights-based approach rooted in EU and international law. This framework outlines four key ethical principles: human autonomy, harm prevention, fairness, and explicability. Chapter II of the HLEG guidelines then translates these abstract principles into seven practical requirements for implementation throughout the AI lifecycle, suggesting technical and non-technical methods. Chapter II is the main section of the HLEG Guidelines and, as said before, strongly influenced the AI Act. In fact, the seven requirements are included in Recital 27 of the AI Act, are part of the voluntary codes of conduct, and have been used in some of the substantive regulations' articles.

"Assessing Trustworthy AI," Chapter III of the HLEG Guidelines, transforms the key requirements for Trustworthy AI into concrete steps through the "Assessment List for Trustworthy AI" (ALTAI). This assessment list provides actionable questions for AI developers and has indirectly informed the AI Act's high-risk AI system requirements. ALTAI was refined through stakeholder feedback during a piloting process, and this feedback was also taken into account by the European Commission in preparing the AI Act.

Essentially, the HLEG Guidelines have a threefold impact on the AI Act:

- they shaped the regulatory requirements for various risk levels of AI;
- they offer ethical direction for voluntary codes of practice and
- they provide overarching ethical principles for all AI practitioners, irrespective of the system's risk classification.

[4] https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

Despite the importance of ALTAI and the AI Act, which is mandatory because it is a legal regulation, other guidelines are worth mentioning to provide a complete background to this deliverable. Thus, the following sub-section will briefly describe some of the most relevant ones.

2.2.2 International and Industry AI Ethical Guidelines: Key Examples

Recognizing a growing need for ethical AI standards, the European Commission formed the HLEG partly in response to individual EU member states' plans for AI ethics initiatives. This was aimed at preventing a fragmented approach within the European market. Simultaneously, a global trend emerged, with countries and international organizations like Singapore[5], Japan[6], and the OECD[7], developing their own AI governance frameworks. The OECD's principles played a significant role in shaping the G20's AI guidelines[8]. Despite criticism regarding their non-binding nature, these diverse efforts were viewed as essential for establishing a shared understanding of ethical AI development.

Another important guideline is the UNESCO Recommendation on the Ethics of Artificial Intelligence[9]. The recommendation incorporates the following principles:

1. Proportionality and do no harm.
2. Safety and Security.
3. Fairness and no discrimination.
4. Sustainability.
5. Right to privacy and data protection.
6. Human oversight and determination.
7. Transparency and explainability.
8. Responsibility and accountability.
9. Awareness and literacy.

As can be seen, these principles are similar to the Trustworthy AI approach followed by the European Union.

As for the private sector, some big companies have also designed their guidelines. For example, Telefónica (Spain) issued a five principles guide for responsible AI titled "Responsible AI by Design in Practice."[10] This guide was inspired by the recommendations of the Berkman Klein Center for Internet & Society at Harvard University[11]. The five principles of Telefonica's guide are:

1. Fair AI: avoid results that end in discrimination
2. Transparent and Explainable AI: the end user must know when she is interacting with an AI

[5] https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework
[6] https://www.jdla.org/en/en-document/en-ai-governance-eco-system/
[7] https://www.oecd.org/en/topics/ai-principles.html
[8] https://oecd.ai/en/wonk/documents/g20-ai-principles
[9] https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence
[10] https://telefonicatech.com/en/blog/responsible-ai-by-design-in-practice
[11] https://cyber.harvard.edu/topics/ethics-and-governance-ai

3. Human-Centered AI: AI must be aligned with the UN Sustainable Development Goals
4. Privacy and Security by Design
5. Extension of these principles to any third party.

2.3 AI Act

The European Commission's AI Act regulates AI development in the EU by focusing on trustworthiness defined by risk levels. The Act employs a risk-based categorization system, where regulations vary based on the potential harm of AI systems. This ensures that safety and ethical considerations are addressed alongside innovation, establishing the EU as a leader in human-centric AI.

The AI Act defines an 'AI system' in Article 3(1) as a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, **i**nfers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.

The AI Act introduces a risk-based classification system to ensure that AI technologies are developed and deployed in a way that prioritizes safety and fundamental rights. This system categorizes AI systems into four different risk levels, each with corresponding regulatory requirements:

Unacceptable Risk: AI systems in this category pose such a significant threat to fundamental rights and public safety that they are banned from being used within the EU. The EU AI Act mentions several examples of such systems:

a. **Social Scoring**: AI that rates individuals based on behavior or characteristics, leading to potential discrimination,

b. **Trait-Based Crime Prediction**: AI predicting future crimes based on personal traits,

c. **Biometric ID in Public**: Real-time biometric identification in public spaces, with limited law enforcement exceptions,

d. **Subliminal Manipulation**: AI that alters behavior without user awareness, potentially causing harm,

e. **Biometric Categorisation**: Categorising people based on sensitive traits like ethnicity or religion,

f. **Emotion Recognition**: AI assessing emotions in workplaces or schools, allowed only for safety purposes,

g. **Facial Image Scraping**: Creating databases by scraping facial images from the web,

h. **Exploitation of Vulnerabilities**: AI exploiting age, disability, or social status to cause harm.

High-risk AI systems: These are AI technologies that, while presenting potential threats to individuals' safety or fundamental rights—especially in critical areas such as healthcare, education, or law enforcement—are not banned. Instead, they are subject to strict regulatory requirements to ensure they operate safely and ethically. This includes performing rigorous conformity assessments before deployment (such as FRIA, Fundamental Rights Impact Assessment[12]) ensuring the system meets established standards.

Limited-risk AI systems: These AI applications, like chatbots or AI-generated content, are subject to lighter regulations than high-risk AI systems and primarily have transparency obligations. Providers must inform users that they are interacting with AI, ensuring users are aware and can make informed decisions. While the regulatory requirements are minimal, these systems must maintain ethical standards and avoid misleading users.

Minimal risk AI systems: these AI systems pose the least threat to users and society, and examples are AI in video games or spam filters. These systems have no mandatory regulatory requirements, but developers are encouraged to follow voluntary codes of conduct to ensure ethical use. The focus is on promoting transparency, fairness, and safety, despite minimal regulatory burden.

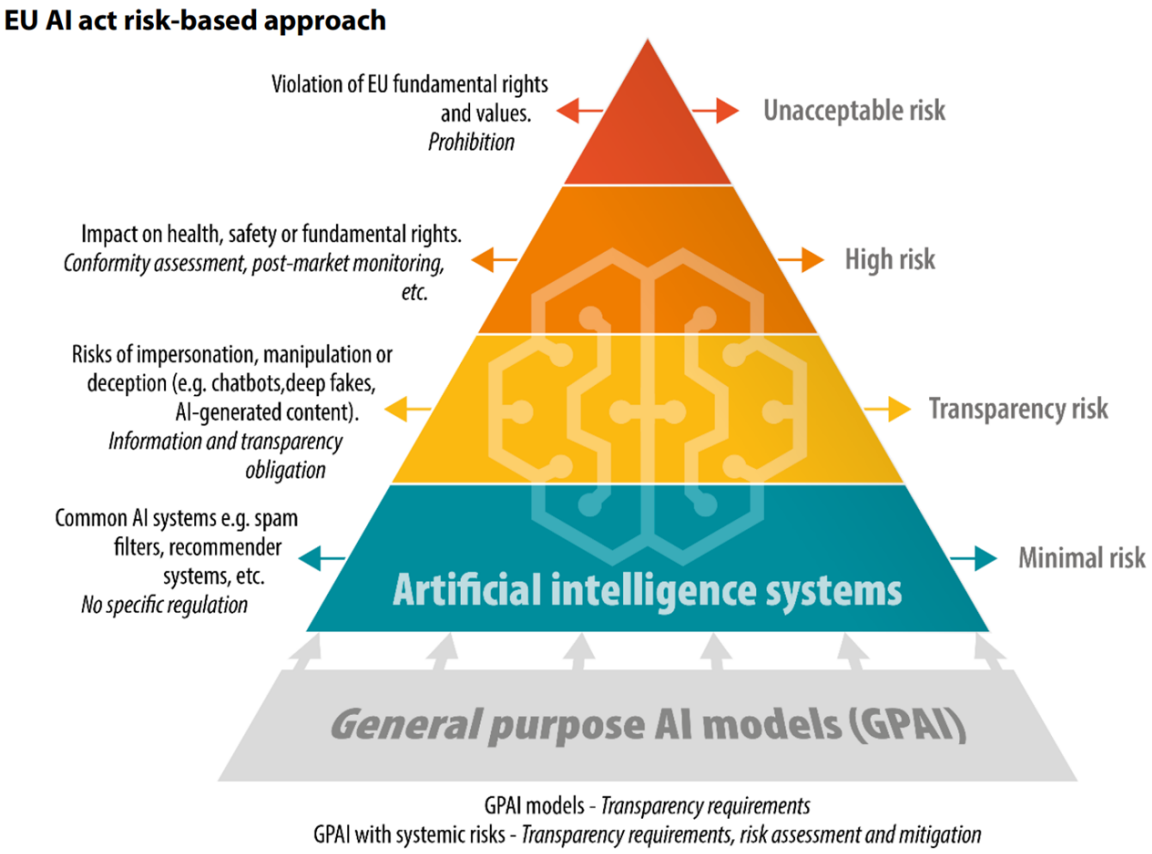

Figure 3: AI-Act risk-based approach

[12] FRIA is a risk management tool. It consists an assessment of the potential impact of an AI system on the rights of any individual that might be affected by the operation of this system.

2.4 Why Funding Bodies Play a Key Role in Trustworthy AI

A key strategy to promote and ensure the development and deployment of Trustworthy AI systems is to shape the allocation of public and charitable funding. By integrating ethical considerations into funding processes, we can incentivize AI developers to prioritize responsible and transparent design practices (Gardner et al, 2022).

Numerous examples of funding initiatives already incorporate ethical requirements and additional monitoring measures. For instance, value for money is a common criterion in funding evaluations, demonstrating that ethical and accountability standards are not new for funders. Therefore, adapting funding procedures to include Trustworthy AI principles should not necessitate a drastic shift in mindset or an entirely new set of requirements. Instead, it can be seen as an extension of existing due diligence practices, ensuring that AI projects align with ethical standards and societal expectations.

Targeting the funding process for AI systems could serve as a powerful lever for change, encouraging applicants to educate themselves on and implement Trustworthy AI principles. By making ethical AI development a prerequisite for funding, funders can drive a broader cultural shift in AI research and deployment, reinforcing the importance of transparency, fairness, and accountability in AI-driven innovations.

This deliverable is intended to provide tools to assess the "Trustworthy AI Statement" in grant proposals; we assume all call for applications will include this statement as a mandatory step. This statement would require applicants to explicitly outline their steps to ensure their AI system adheres to ethical guidelines and can be evaluated against rigorous standards. By embedding this requirement into funding mechanisms, we introduce a subtle yet effective behavioral nudge encouraging AI developers to incorporate ethical design from the outset.

Even minor adjustments within the AI funding ecosystem can initiate meaningful change, helping to mitigate the ethical risks highlighted in the AI4Gov project activities. By leveraging funding as a tool for trustworthy AI use and development, we can foster a landscape where only AI systems that meet high ethical and trustworthiness standards receive financial support, ultimately benefiting society and technological progress. To that end, the following chapter will offer a set of criteria, parameters, and a checklist for funding bodies to assess the applicants' Trustworthy AI Statement.

3 Report on Criteria for Funding Bodies Evaluation of Trustworthy AI Statement

When assessing AI projects for funding, it is essential to distinguish between parameters and criteria:

- Parameters define broad areas of evaluation that guide the assessment.
- Criteria are specific, measurable elements within each parameter that determine the project's level of trustworthiness.

This document outlines key parameters and their corresponding criteria based on the legal, ethical, and technical robustness pillars.

3.1 Parameters and Criteria

1. Human Agency and Oversight

- **Parameter:** Ensuring human control and oversight in AI decision-making.
- **Criteria**:
    - Existence of mechanisms for human intervention and override.
    - AI system enhances, rather than replaces, human decision-making.
    - Users understand the AI's role and capabilities.

2. Technical Robustness and Safety

- **Parameter**: Ensuring AI operates reliably and safely in diverse conditions.
- **Criteria**:
    - AI system undergoes rigorous testing for Security and reliability.
    - Mechanisms exist to prevent adversarial attacks and system failures.
    - System includes fallback mechanisms in case of errors.

3. Privacy and Data Governance

- **Parameter:** Ensuring responsible data management and user privacy.
- **Criteria:**
    - Compliance with privacy laws and data protection regulations.
    - Implementation of encryption and anonymization techniques.
    - Clear policies on data collection, storage, and user consent.

4. Transparency

- **Parameter:** Ensuring clear communication of AI system functionalities.

- **Criteria:**
    - Availability of understandable documentation for users and stakeholders.
    - Ability for external audits and review of system decisions.
    - Explainability of AI model outputs.

5. Diversity, Non-discrimination, and Fairness

- **Parameter:** Ensuring AI does not reinforce biases or discrimination.
- **Criteria:**
    - Assessment of potential biases in datasets and models.
    - Implementation of fairness checks across demographic groups.
    - Inclusion of diverse stakeholders in AI system development.

6. Societal and Environmental Well-being

- **Parameter:** Ensuring AI contributes positively to society and the environment.
- **Criteria:**
    - Evaluation of social impact and ethical considerations.
    - Energy-efficient development and deployment practices.
    - Alignment with human-centered values and sustainability goals.

7. Accountability

- **Parameter:** Clearly defining responsibility and compliance mechanisms.
- **Criteria:**
    - Assignment of accountability for AI decisions and errors.
    - Existence of redress mechanisms for those affected by AI decisions.
    - Compliance with relevant legal and ethical standards.

Using these parameters and criteria, evaluators can ensure that AI projects meet the standards for trustworthiness and responsible AI development.

3.2 Trustworthy AI Project Evaluation Checklist

Instructions: Evaluate each criterion based on the provided scoring system. The total score will indicate the project's alignment with Trustworthy AI principles.

Scoring System:

- **5:** Fully meets the requirement
- **4:** Mostly meets the requirement

- **3:** Partially meets the requirement
- **2:** Minimally meets the requirement
- **1:** Does not meet the requirement

Legal, Ethical, and Technical Robustness Pillars

1. **Human Agency and Oversight**
    - Ensures human oversight in decision-making ( )
    - Enables meaningful human control over AI outputs ( )
    - Provides mechanisms for human intervention or override ( ) **Subtotal: /15**
2. **Technical Robustness and Safety**
    - Demonstrates reliability and Security in various conditions ( )
    - Includes mechanisms for robustness against adversarial attacks ( )
    - Plans for system resilience and fallback mechanisms ( ) **Subtotal: /15**
3. **Privacy and Data Governance**
    - Implements strong data protection measures ( )
    - Ensures compliance with relevant privacy laws ( )
    - Adopts secure data governance and anonymization practices ( ) **Subtotal: /15**
4. **Transparency**
    - Provides clear explanations of AI system functioning ( )
    - Offers understandable documentation for stakeholders ( )
    - Allows external auditing and review of system decisions ( ) **Subtotal: /15**
5. **Diversity, Non-discrimination, and Fairness**
    - Assesses and mitigates potential biases in AI models ( )
    - Ensures fair treatment across different demographic groups ( )
    - Engages diverse stakeholders in development and evaluation ( ) **Subtotal: /15**
6. **Societal and Environmental Well-being**
    - Evaluates the system's impact on society ( )
    - Supports environmental sustainability and energy efficiency ( )
    - Aligns with ethical and human-centered values ( ) **Subtotal: /15**

7. **Accountability**
    - Clearly assigns responsibility for AI decisions and outcomes ( )
    - Includes mechanisms for redress and recourse ( )
    - Ensures compliance with legal and ethical standards ( ) **Subtotal: /15**

TOTAL SCORE: /105

Evaluation Key:

- **90-105:** Highly trustworthy AI project
- **75-89:** Meets most requirements but needs improvement
- **60-74:** Partially meets requirements; significant improvements needed
- **Below 60:** Does not meet minimum requirements

4 Self-assessment tools

4.1 Statement of Support (SoS) (for funding bodies)

The SoS is a valuable tool that can support funding bodies and other organizations in evaluating AI-related proposals. Each funding body's role when financing AI-based projects is essential to ensure that such developments are in the human interest, respect fundamental rights, and do not cause environmental damage. The advances of AI-based systems are too fast to be captured and contained solely by legal regulation; on the other hand, it is undesirable for legislation to operate as an unnecessary brake on an advance that could mean remarkable improvements for humanity. In this sense, the decision of which projects to fund should be guided by their adherence to globally accepted ethical principles on AI and compliance with current legal regulations. On top of evaluating the applicants' proposals' technical strengths, reviewing their Trustworthy AI Statements is critical. This document will help you to do so by providing guidance on which areas should be considered when assessing those statements.

4.1.1 Instructions

This document is a Statement of Support (SoS) for you to use while evaluating Trustworthy AI Statements from applicants seeking funding to develop AI-based systems. The SoS was developed in conjunction with a self-assessment tool for applicants called Stop-and-Think, which sets out the key areas they should review to ensure compliance with ethical and legal standards[13]. As such, the SoS reflects the same areas of assessment, and we recommend that, where possible, applicants be offered Stop-and-Think to make the application of the SoS more efficient. As the SoS encompasses a selection of mandatory and complimentary requirements and principles, it is advisable to be aware of this difference as it is clarified in each step. At the same time, because a call for application may include different kinds of AI systems, for example, high-risk or limited-risk, this tool is better understood as a holistic approach to evaluating applications. However, funding bodies could also adapt and convert this tool into a scoring tool if it is considered more appropriate.

The SoS has two uses. First, it can be used as a guide on key aspects of deciding whether each project under consideration meets ethical and legal requirements. Second, the SoS allows projects to be compared to decide which are more appropriate according to the ethical and legal parameters evaluated.

4.1.2 Key Areas to Consider when assessing Trustworthy AI Statements

Step 1: Did the applicant correctly classify the AI system proposed following the AI Act Risk Classification?

First, a brief summary of the AI Act classification. The EU AI Act classifies AI systems into four risk categories: Unacceptable Risk, High Risk, Limited Risk, and Minimal Risk.

[13] Listed in the Annex of this tool.

- **Unacceptable Risk**: AI systems that deploy harmful manipulative "subliminal techniques"; AI systems that exploit specific vulnerable groups (physical or mental disability); AI systems used by public authorities or on their behalf, for social scoring purposes, "Real-time" remote biometric identification systems in publicly accessible spaces for law enforcement purposes, except in a limited number of cases.
- **High Risk**: AI systems that adversely impact people's safety or fundamental rights. The AI Act differentiates between two categories of high-risk systems. Systems used as a safety component of a product falling under EU health and safety harmonization legislation; systems deployed in eight specific areas detailed in Annex III.
- **Limited Risk**: AI systems that interact with humans (e.g., chatbots), emotion recognition systems, biometric categorization systems, and AI systems that generate or manipulate image, audio, or video content (e.g., deepfakes) would be subject to a limited set of transparency obligations.
- **Minimal Risk**: these systems could be developed and used without conforming to any additional requirements.

If the applicant correctly determines the risk level, the first step in the analysis will result in a positive result. If the applicant fails to do so, you should pay extra attention to the rest of the steps. In addition, it is key to note that prohibited practices should not be funded.

Step 2: Application for developing a High-Risk AI system

If the AI system is classified as high-risk, by you or the applicant, ensure it complies, among others, with the following requirements:

- **Risk Management System**: Implement a risk management system to identify, assess, and mitigate risks.
- **Data Governance**: Ensure the quality and integrity of the data used. This includes proper data collection, annotation, and handling procedures.
- **Technical Documentation**: Maintain comprehensive technical documentation detailing the system's purpose, design, development, testing, and deployment.
- **Record Keeping**: Create a system that allows automatic recording of events (logs) over the lifetime of the system.
- **Transparency and Information Provision**: Provide clear information to users about the system's capabilities and limitations.
- **Human Oversight**: Design mechanisms that allow human oversight and intervention when necessary.
- **Robustness, Accuracy, and Security**: Ensure your system is resilient, accurate, and secure against potential threats.

While it is not mandatory, implementing the safeguards required for high-risk AI systems in non-high-risk AI systems can be considered best practice. Therefore, if the applicant of a non-high-risk AI system implements these safeguards, his or her application should be granted extra weight when compared to others.

Step 3: Ethical Considerations

Adhering to ethical principles is critical to complying with guidelines, the AI Act, ALTAI, and other instruments on which this tool is based. Ensure the Trustworthy AI Statement discusses potential ethical challenges such as biases, misuses, unintended harms, impact on equality, and proportionality between the proposed system and the intended goals. **Remember, ethical considerations go beyond what is legally mandatory. Something can be legal but unethical or illegal but ethical. The applicant should explain how and why their proposal is ethically aligned in the Trustworthy AI Statement.** The following fundamental principles should be present and developed in the project, and the applicant should state them in their AI Trustworthy Statement:

- **Human agency and oversight:** AI systems should empower human beings and foster their fundamental rights and should be subject to proper oversight mechanisms
- **Technical robustness and safety:** To avoid unintentional harm, AI systems should be resilient, secure, accurate, reliable and reproducible.
- **Privacy and data governance**: Comply with GDPR and other relevant privacy regulations. Ensure the AI system does not infringe on individuals' privacy rights.
- **Transparency**: the data, system and AI business models should be transparent and individuals need to know they are interacting with an AI system. The decisions taken by an AI system should be explained and easily understandable for the individual concerned.
- **Non-discrimination and fairness**: Design your AI system to avoid bias and discrimination. Implement measures to detect and mitigate any potential bias in data and algorithms.
- **Societal and environmental well-being**: AI systems should benefit all human beings, including future generations. It must hence be ensured that they are sustainable and environmentally friendly.
- **Accountability**: Establish clear responsibility for the AI system's decisions and actions. Ensure processes are in place for redress and remedy in case of harm or misuse.

Step 4: Transparency and User Awareness

Transparency and user awareness are key to trustworthy AI. The statement you are evaluating should be robust in this regard, ***regardless of the risk classification***. Revise the Trustworthy AI Statement to be sure that it complies with the standards put forward by the AI Act:

- **Clear Communication**: Inform individuals when they are interacting with an AI system. Provide understandable information about how the AI system makes decisions. For instance, provide individuals with information about
  - When AI technologies are being used;
  - The capabilities and limitations of a given model;
  - The data on which the model was trained;
  - The data used to generate outputs;
  - Whether data is retained (and if so, what and for how long);
  - Avenues to remediate or appeal outputs produced by the model; and
  - Whether user choices can influence system performance.
- **Documentation for Users**: Offer comprehensive documentation and user guides that explain the AI system's functionality, limitations, and correct usage.

Step 5: Sustainability and Societal Impact

Finally, review how the Trustworthy AI Statement refers to proposed project's sustainability and societal impact. In addition, **if you are assessing similar proposals, take this dimension into special consideration to decide which proposal is the best.** For example, you should consider if some of the following environmental challenges and discrimination problems are analyzed in the statement:

- **Environmental Impact**: Aims for energy-efficient algorithms and sustainable practices.
- **Social Impact**: Evaluates the broader societal implications of the AI system. Ensures it contributes positively to society and does not reinforce existing inequalities or create new ones.

Final Consideration: Trustworthy AI Statement Checklist

Before making a decision, ensure you have addressed the following:

1. **Risk Assessment**: Did the applicant correctly determine the risk level? If it is a High-Risk level system, is it proportionate to the goals it has?
2. **Ethical Considerations**: Has the Trustworthy AI Statement went through all the ethical considerations and explained how the applicant will meet them?
3. **Transparency Measures**: If the project is funded, is the applicant ready to provide users with clear information and documentation?

4. **Impact Assessment**: If the project is funded, is the applicant ready to evaluate and mitigate the environmental and societal impacts of your AI system?

4.1.3 Annex

- EU: AI Act and EC ethical guidelines
- AI Impact Assessment. A tool to set up responsible AI projects, Ministry of Infrastructure and Water Management
- Framework Convention on AI
- USA: Blueprint for AI Bill of Rights
- AU AI ethical principles
- CA Responsible use of AI
- IEEE ethically aligned design
- Human Rights, Democracy, and the Rule of Law Assurance Framework for AI Systems: A Proposal
- Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment
- EU model contractual AI clauses to pilot in procurements of AI
- IEEE CertifAIEd™ – Ontological - Specification for Ethical Algorithmic Bias
- Human Rights, Democracy, and the Rule of Law Assurance Framework for AI Systems: A Proposal
- CAN/CIOSC 101:2019 - Ethical design and use of automated decision systems: AI Act draft and EC ethical guidelines
- Center for Inclusive Change, Essential Considerations in AI Contracting
- WEF, Guidelines for AI procurement
- WEF, AI Procurement in a Box: AI Government Procurement Guidelines
- Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems (Canada)
- ISO/IEC 42001:2023(en) Information technology — Artificial intelligence — Management system

4.2 Stop-and-Think Self-Assessment tool for applicants

The following self-assessment tool will help you complete your Trustworthy AI Statement thoroughly and efficiently. Your role as a potential recipient of funds to develop AI-based systems is critical in respecting globally agreed ethical principles. Adverse effects on human rights and the AI environment are a reality. One way to avoid them is to evaluate the project and present a thorough analysis before implementation. You must understand that the Trustworthy AI Statement does not simply state that you will comply with mandatory legal regulations but implies a solid commitment to the ethical development of AI-based systems. This instrument particularly considers the AI Act, the ALTAI, and tools developed by different governments from EU Member States to evaluate the impact of AI systems[14]. Compliance with ethical and legal standards may require the assistance of experts in these areas. Consider consulting with such experts if the system you intend to develop involves human rights impacts.

4.2.1 Instructions

The **Stop-and-Think** is a guide to be used during the ethical reflection on the proposed AI-based system development and during the drafting of the **Trustworthy AI Statement**. The Stop-and-Think tool will guide you step by step to elaborate on how you comply with the critical ethical and legal areas and not simply state that you will abide by them. Please note that this tool does not preclude you from adding further specifications depending on the type of project for which you are applying for funding.

4.2.2 Key areas to consider when preparing your Trustworthy AI Application Statement

Stage 1: Understanding AI Act Risk Classification

Stop: Before preparing your statement, ensure all your group members understand the AI Act Risk Classification.

Think: Read carefully the summary of the AI Act classification and decide according to it.

1. Risk Classification

- **Understand Risk Levels**: The EU AI Act classifies AI systems into four risk categories: Unacceptable Risk, High Risk, Limited Risk, and Minimal Risk.
- **Unacceptable Risk**: AI systems that deploy harmful manipulative "subliminal techniques"; AI systems that exploit specific vulnerable groups (physical or mental disability); AI systems used by public authorities or on their behalf, for social scoring purposes, "Real-time" remote biometric identification systems in publicly accessible spaces for law enforcement purposes, except in a limited number of cases.

[14] Listed in the Annex of this tool.

- **High Risk**: AI systems that adversely impact people's safety or fundamental rights. The AI Act differentiates between two categories of high-risk systems. Systems used as a safety component of a product falling under EU health and safety harmonization legislation; systems deployed in eight specific areas detailed in Annex III.
- **Limited Risk**: AI systems that interact with humans (e.g., chatbots), emotion recognition systems, biometric categorization systems, and AI systems that generate or manipulate image, audio, or video content (e.g., deepfakes) would be subject to a limited set of transparency obligations.
- **Minimal Risk**: these systems could be developed and used without conforming to any additional requirements.

Hint: use the AI Checker https://artificialintelligenceact.eu/assessment/eu-ai-act-compliance-checker/

Stage 2: Identify Your System's Risk Level

Stop: Which classifications mentioned above does your AI system project fall under?

Think: Determine which category your AI system falls into. Remember that high-risk systems include those used in critical infrastructure, education, employment, essential public services, law enforcement, and migration, among others. **In addition, it is key to note that prohibited practices will not be funded.**

Stage 3: Application that includes a High-Risk AI system [for high-risk AI Systems]

Stop: Revise the High-Risk AI System Requirements of the AI Act.

Think: If your AI system is classified as high-risk, ensure it complies, among others, with the following requirements:

- **Risk Management System**: Implement a risk management system to identify, assess, and mitigate risks.
- **Data Governance**: Ensure the quality and integrity of the data used. This includes proper data collection, annotation, and handling procedures.
- **Technical Documentation**: Maintain comprehensive technical documentation detailing the system's purpose, design, development, testing, and deployment.
- **Record Keeping**: Create a system that allows automatic recording of events (logs) over the lifetime of the system.
- **Transparency and Information Provision**: Provide clear information to users about the system's capabilities and limitations.

- **Human Oversight**: Design mechanisms that allow human oversight and intervention when necessary.
- **Robustness, Accuracy, and Security**: Ensure your system is resilient, accurate, and secure against potential threats.

Hint: While it is not mandatory, implementing the safeguards required for high-risk AI systems in non-high-risk AI systems can be considered best practice.

Stage 4: Ethical Considerations

Stop: Adhering to ethical principles is critical to complying with guidelines, the AI Act, ALTAI, and guidelines developed by governments and other organisms on which this instrument is based. Ensure your team discusses potential ethical challenges such as biases, misuses, unintended harms, impact on equality, and proportionality between the proposed system and the intended goals. **Remember, ethical considerations go beyond what is legally mandatory. Something can be legal but unethical or illegal but ethical. In your Trustworthy AI Statement, you should explain how and why your proposal is ethically aligned**.

- **Think:** Revise the following fundamental principles when developing your project and AI Trustworthy Statement. **Human agency and oversight:** AI systems should empower human beings and foster their fundamental rights and should be subject to proper oversight mechanisms
- **Technical robustness and safety:** To avoid unintentional harm, AI systems should be resilient, secure, accurate, reliable and reproducible.
- **Privacy and data governance**: Comply with GDPR and other relevant privacy regulations. Ensure the AI system does not infringe on individuals' privacy rights.
- **Transparency**: the data, system and AI business models should be transparent and individuals need to know they are interacting with an AI system. The decisions taken by an AI systems should be explained and easily understandable for the individual concerned.
- **Non-discrimination and fairness**: Design your AI system to avoid bias and discrimination. Implement measures to detect and mitigate any potential bias in data and algorithms. .
- **Societal and environmental well-being**: AI systems should benefit all human beings, including future generations. It must hence be ensured that they are sustainable and environmentally friendly.
- **Accountability**: Establish clear responsibility for the AI system's decisions and actions. Ensure processes are in place for redress and remedy in case of harm or misuse.

Stage 5: Transparency and User Awareness

Stop: Before fully developing the AI system, ensure transparency and user awareness are considered and that AI Act standards are followed, ***regardless of the risk classification.***

Think: Does the project include the following critical points?

- **Clear Communication**: Inform individuals when they are interacting with an AI system. Provide understandable information about how the AI system makes decisions. For instance, provide individuals with information about
  - When AI technologies are being used;
  - The capabilities and limitations of a given model;
  - The data on which the model was trained;
  - The data used to generate outputs;
  - Whether data is retained (and if so, what and for how long);
  - Avenues to remediate or appeal outputs produced by the model; and
  - Whether user choices can influence system performance.
- **Documentation for Users**: Offer comprehensive documentation and user guides that explain the AI system's functionality, limitations, and correct usage.

Stage 6: Sustainability and Societal Impact

Stop: Before delving deeper into the technical aspects of AI system development, stop and reflect on the proposed project's sustainability and societal impact.

Think: Consider the following environmental challenges and discrimination problems that may arise from your project.

- **Environmental Impact**: Consider the environmental impact of developing and deploying your AI system. Aim for energy-efficient algorithms and sustainable practices.
- **Social Impact**: Evaluate the broader societal implications of your AI system. Ensure it contributes positively to society and does not reinforce existing inequalities or create new ones.

Final Consideration: Trustworthy AI Statement Checklist

Before submitting your funding application, ensure you have addressed the following:

1. **Risk Assessment**: Have you classified your AI system's risk level?
2. **Ethical Considerations**: Have you implemented measures to ensure fairness, privacy, and accountability?
3. **Transparency Measures**: If your project is funded, are you ready to provide users with clear information and documentation?

4. **Impact Assessment**: If your project is funded, are you ready to evaluate and mitigate the environmental and societal impacts of your AI system?

4.2.3 Annex

- EU: AI Act and EC ethical guidelines
- AI Impact Assessment. A tool to set up responsible AI projects, Ministry of Infrastructure and Water Management
- Framework Convention on AI
- USA: Blueprint for AI Bill of Rights
- AU AI ethical principles
- CA Responsible use of AI
- IEEE ethically aligned design
- Human Rights, Democracy, and the Rule of Law Assurance Framework for AI Systems: A Proposal
- Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment
- EU model contractual AI clauses to pilot in procurements of AI
- IEEE CertifAIEd™ – Ontological - Specification for Ethical Algorithmic Bias
- Human Rights, Democracy, and the Rule of Law Assurance Framework for AI Systems: A Proposal
- CAN/CIOSC 101:2019 - Ethical design and use of automated decision systems: AI Act draft and EC ethical guidelines
- Center for Inclusive Change, Essential Considerations in AI Contracting
- WEF, Guidelines for AI procurement
- WEF, AI Procurement in a Box: AI Government Procurement Guidelines
- Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems (Canada)
- ISO/IEC 42001:2023(en) Information technology — Artificial intelligence — Management system

5 Conclusions

This document presents tools for public, private and other types of organizations that fund projects, systems or services that use AI and have to take into account ethical and legal aspects. These requirements are essential in order to ensure that the outcomes do not create any harm to people, are compliant with the EU legislation, foster public trust and acceptance, and assure the accountability of this type of organizations. A set of tools is produced for this purpose and are presented in this deliverable.

6 References

COECKELBERGH, M., “Why AI Undermines Democracy”, Polity Press, 2024

COECKELBERGH, M., “AI Ethics”, The MIT Press, 2020

DÍAZ-RODRÍGUEZ, Natalia et al., *Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation*, Information Fusion, Volume 99, 2023, 101896, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2023.101896. (https://www.sciencedirect.com/science/article/pii/S1566253523002129)

KLUGE CORRÊA, Nicholas, et al., *Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance*, Patterns, Volume 4, Issue 10, 2023, 100857, ISSN 2666-3899, https://doi.org/10.1016/j.patter.2023.100857, (https://www.sciencedirect.com/science/article/pii/S2666389923002416)

SMUHA, Nathalie A., The Work of the High-Level Expert Group on AI as the Precursor of the AI Act (October 01, 2024). Available at SSRN: https://ssrn.com/abstract=5012626 or http://dx.doi.org/10.2139/ssrn.5012626

GARDNER, A., SMITH, A.L., STEVENTON, A. *et al.* Ethical funding for trustworthy AI: proposals to address the responsibilities of funders to ensure that projects adhere to trustworthy AI practice. *AI Ethics* **2**, 277–291 (2022). https://doi.org/10.1007/s43681-021-00069-w

ULINCANE, “Artificial Intelligence in the European Union”, in “The Routledge Handbook of European Integrations”, 2022.