D2.1

FAITH Methodological Framework and Requirements Analysis v1

Related Work Package	**WP2 – FAITH AI Framework for trustworthiness & coordination among Large Scale Pilots (LSPs)**
Related Task	Task 2.1. - FAITH Trustworthiness Assessment Framework (FAITH_TAF)
Lead Beneficiary	trustilio
Contributing Beneficiaries	All partners involved in Tasks: 2.1
Document version	v.9.0
Deliverable Type	R-Document, Report
Distribution level	PU-Public
Contractual Date of Delivery	31/03/2025
Actual Date of Delivery	31/03/2025

Contributors	Theofanis Fotis (trustilio) Kitty Kioskli (trustilio) Eleni Seralidou (trustilio) Nineta Polemi (trustilio) Vasiliki Antoniou (trustilio) Abdullah Elbi (KU Leuven) Jean De Meyere (KU Leuven)

	Sara Garsia (KU Leuven) Ana Maria Corrêa (KU Leuven) Andrea Carboni (CNR) Sara Colantonio (CNR) Silvia Gravili (CNR) Giuseppe Riccardo Leone (CNR) Davide Moroni (CNR) Maria Antonietta Pascali (CNR) D.I Fotiadis (FORTH) M. Tsiknakis (FORTH) N. Tachos (FORTH) V. Pezoulas (FORTH) D. Zaridis (FORTH) G. Kaliatakis (FORTH) Asbjørn Følstad (SINTEF) G. Mentzas (ICCS) D. Apostolou (ICCS) E. Anagnostopoulou (ICCS) N. Grammatikos (ICCS) E. Tsalapati (ATC) S. Modafferi (UoS)
Reviewers	Jean De Meyere (KUL) Pilar Sala (AOA)

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 Directorate-General for Communications Networks, Content and Technology. Neither the European Union nor the granting authority can be held responsible for them.

Version history

Version	Description	Date completed	Contributors
	ToC	21.2.2024	trustilio
	Finalised Toc	27.4.2024	trustilio and all
V1	Contributions	24.5.24	trustilio
V2	Contributions Check point 1	30.06.2024	trustilio
V3	Contributions Check point 2	30.09.2024	trustilio
V4	Contributions/review	29/10/24	all
V5	Contributions Check point 3	30.11.2024	trustilio
V6	Contributions Final check point	28.02.2025	trustilio
	Completion of peer review	09.03.2025	
V7	Address reviewers' major comments	16.03.2025	trustilio
V8	Address minor comments	23.03.2025	trustilio
V9	D2.1 Submission	31.03.2025	PC

Statement of Originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgment of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Disclaimer

This document contains material, which is the copyright of one or more FAITH consortium parties, and may not be reproduced or copied without permission.

All FAITH consortium parties have agreed to full publication of this document.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the FAITH consortium as a whole, nor individual FAITH consortium parties, warrant that the information contained in this document is capable of use, nor that use of the information is free from risk, accepting no liability for loss or damage suffered by any person using this information.

Executive summary

This deliverable reflects the first outcomes of T2.1 - FAITH Trustworthiness Assessment Framework (FAITH_TAF) (M1-M24) [Lead: trustilio, Partners: UoS, SINTEF, ICCS, CNR, AOA, FORTH].

The main outcomes of T2.1 that are described in this deliverable include:

An updated state of the art analysis on technological, policy, standards and legal aspects of AI trustworthiness. The first version of the FAITH AI Trustworthiness Assessment Framework (FAITH_TAF) has been proposed in this deliverable considering the NIST AI RFM, ENISA's AI cybersecurity recommendations, and EU legal instruments as well as the ISO2700x series of standards. The task also identified and measured the cognitive, psychological, social, behavioural characteristics, and vulnerabilities of human teams involved in the AI lifecycle that may affect AI trustworthiness and trust perception.

The methodology adopted in this Task and has been described here is based upon:

a) Analysing the trustworthiness of AI systems, focusing on fairness, technical accuracy and robustness, the socio-technical environment of the AI, user perceptions, and EU ethical and democratic principles.

b) Conducting research on a risk assessment-based approach to evaluating and optimizing AI trustworthiness suitable for the EU context.

c) Identifying suitable AI technologies and resources to achieve trustworthiness, such as AI/ML modelling, decision intelligence, serious games, anomaly detection, rules-based knowledge management, and anonymous human profiling.

d) Determining the ethical and legal requirements for FAITH outcomes based on current and upcoming regulatory instruments and interpretations, assessing their applicability in the FAITH context, identifying gaps, and making recommendations to address them.

e) Proposing psychosocial profiles with human traits that determine the trustworthiness of the AI participants in AI-lifecycle and measure the degree of their trustworthiness.

f) Developing metrics and scales for measuring AI risks and trustworthiness, which have been proposed and documented in D.2.1.

g) Estimating the risks for trustworthiness of an AI system considering not only the technical but also the social and human threats.

h) Selecting measurements/controls for managing trustworthiness risks that are technical, social, behavioural, legal and policy related.

The proposed FAITH AI_TAF and measurements proposed here (in D.2.1) will be further validated via workshops with domain experts, users, AI participants from pilot domains, and affected communities and finalized in D.2.2. The first version of the AI team maturity measurements proposed have already validated in the 1st FAITH workshop (see Annex B). The consequent D2.2 will finalize the work in T.2.1.

Table of Contents

List of Abbreviations

Abbreviation	Explanation
AI	Artificial intelligence
AI Act	Artificial Intelligence Act
ENISA	European Union Agency for Cybersecurity
ETSI	European Telecommunications Standards Institute
EU	European Union
GDPR	general data protection regulation
ISO	International Organization for Standardization
NIS/NIS 2	EU directives on measures for a high common level of cybersecurity across the EU
NIST	National Institute for Standards and Technology
OECD	Organisation for Economic Co-operation and Development
CEN-CENELEC	European Standardisation of Standards

List of Tables

List of Figures

1 Introduction

AI systems present unique **security challenges that differ from traditional software** due to their dynamic, data-driven, and opaque nature. Unlike static codebases, AI models continuously learn from data, making them vulnerable to manipulation through targeted inputs. Additionally, the complexity of AI models makes it challenging to predict their responses to new or adversarial data.

AI systems are also vulnerable to adversarial attacks, where small, subtle changes to input data can cause incorrect outputs. These attacks exploit the fragile nature of AI models, particularly in high-dimensional spaces, and defending against them often requires retraining with adversarial examples or applying defensive techniques.

Another critical issue is biases inherited from training data, leading to unfair outcomes. Such biases can undermine trust and result in legal and reputational damage. Moreover, the broader societal consequence of AI security must be considered, especially for critical infrastructure and essential services. Explainability, another dimension of trustworthiness imposed by the AI Act, is another challenge since **balancing technological innovation and trustworthiness is key**, as revealing and explaining too much about a model's security measures can help adversaries exploit vulnerabilities.

Furthermore, security measures themselves may introduce biases if they disproportionately affect certain demographic groups. Regular audits, bias mitigation, and inclusive design are necessary to prevent these ethical issues. In sum, ethical AI development requires a holistic approach that integrates diverse datasets, continuous audits, adherence to guidelines, and a commitment to fairness and equity, positioning AI security as both a technical challenge and a societal responsibility.

Trustworthiness in Artificial Intelligence (AI) encompasses several dimensions, as outlined by the CEN JTC21, including cybersecurity, transparency, robustness, accuracy, data quality and governance, human oversight, and record keeping. Managing the risks associated with trustworthiness involves identifying, analysing, estimating, and mitigating threats across these dimensions. Implementing a quality management system supports effective risk management, which can be validated through conformity assessment processes.

Additionally, ensuring the reliability of AI systems requires addressing risks related to human elements. Furthermore, ensuring the reliability and effectiveness of AI systems necessitates a comprehensive approach that addresses the risks associated with human factors. The **trustworthy AI maturity of the AI teams within an organization** play a crucial role in shaping the trustworthiness, performance, and ethical standards of AI systems. These human elements, including biases, decision-making processes, and interpersonal dynamics, can significantly influence how AI systems are developed, tested, and deployed. It is essential for organizations to identify, recognize, estimate the AI maturity of their teams; identify human vulnerabilities that can be used to exploit AI threats. Social mitigation actions (e.g. co-creation workshops, behaviour change interventions, awareness and practical trainings) need to be included in the organizations' risk treatment plans. By doing so, organizations can not only enhance the trustworthiness of their AI systems but also foster a culture of trust and reliability in the technologies they create. Human vulnerabilities, such as implicit or explicit biases, can inadvertently influence AI algorithms, leading to biased decision-making outcomes. Additionally, a lack of vigilance in recognizing AI-related threats may result in human errors—one of the most common risks—which can be exploited across the AI lifecycle. These factors introduce vulnerabilities that undermine the reliability and trustworthiness of AI systems. Trust in AI also depends on user comprehension and acceptance, emphasizing the importance of clear communication regarding how AI functions and its limitations. Managing these human-related factors is essential for enhancing AI system reliability and promoting responsible AI development [1].

Furthermore, maintaining the integrity of AI systems involves understanding the interconnectedness of threats across various dimensions. This necessitates a comprehensive evaluation of the effectiveness of controls and mitigation strategies that address all types of potential risks. For instance, challenges related to human oversight—such as biases, transparency deficiencies, and the ability to explain decisions—can compromise cybersecurity and the integrity of data. Conversely, vulnerabilities in cybersecurity can exacerbate the risks associated with human oversight. Effective measures should encompass not only technical solutions but also consider behavioural, social, cultural, and ethical factors. The AI Act, specifically Article 9, mandates a comprehensive approach to risk management that thoroughly evaluates both technical and human-related risks, guiding the current project to implement robust frameworks and methodologies tailored to address these multifaceted challenges effectively.

In this deliverable, we provide **a comprehensive review of current advancements** and ongoing efforts in technology, policy, law, and standardization related to AI trustworthiness. We also present tools designed to assess various dimensions of trustworthiness in AI systems. To further enhance the evaluation process, we introduce the **FAITH_AI_TAF framework**, a structured, step-by-step methodology for identifying and assessing AI threats and vulnerabilities, evaluating their potential consequences, and estimating the associated risks. This framework considers both the criticality of the systems and the maturity of the organization's AI teams. Additionally, we present the **initial design principles of the FAITH System Trust Modeler**, which will implement the FAITH AI_TAF framework, along with user journey examples to illustrate the system's design principles and core functionalities.

1.1 Key Challenges Guiding the Work

It is of high importance to mention that current efforts in AI risk management often neglect human factors and do not introduce metrics for socially or human-related threats.

As noted in the NIST AI RFM [2], further research is necessary to grasp the current limitations of human-AI interaction, a concern also underscored by ENISA, which emphasizes the need for developing ethical and social metrics for AI. Another challenge lies in the adoption and integration of non-technical controls, such as social responsibility, despite existing standards like ISO 26000:2010 and ISO/IEC TR 24368:2022. These standards are not yet fully integrated into AI risk management phases. Effective collaboration among cybersecurity engineers, AI specialists, and professionals from disciplines such as social psychology, behaviour, and ethics is crucial to enhance AI risk management practices, particularly in selecting targeted human-centric controls.

1.2 Scope and Methodology

In this deliverable, we assess and outline existing efforts, initiatives and results in identifying technical, legal, policy, standard related efforts in the AI-trustworthiness. We initiate a framework adopting a risk assessment approach in identifying, estimating and managing trustworthiness risks capturing all its dimensions (cybersecurity, transparency, robustness, accuracy, data quality and governance, human oversight).

The initial version of the FAITH AI Trustworthiness Assessment Framework (FAITH_TAF) is presented in this deliverable, considering the NIST AI RFM, ENISA's AI cybersecurity recommendations, EU legal instruments, the ISO2700x standards, ISO/IEC 5338:2023 since it can be used in the life-cycle stages of the AI system, ISO ISO/IEC TR24028:2020, ISO/IEC 24368:2022 (ethical concerns) and ISO/EC JTC1/SC42 AI initiatives.

This task also identified and measured cognitive, psychological, social, and behavioural characteristics and vulnerabilities of individuals involved in the AI lifecycle that could consequence AI trustworthiness and trust perception.

The first version of the FAITH AI_TAF detailed in this task T2.1 captured in the deliverable D2.1 includes: a) Analysing AI system trustworthiness, focusing on fairness, technical accuracy, robustness, the socio-technical environment, user perceptions, and EU ethical and democratic principles. b) Researching a risk assessment-based approach for evaluating and optimizing AI trustworthiness suitable for the EU context. c) Identifying suitable AI tools, technologies and resources for assessing the various dimensions of trustworthiness, and anonymous human profiling of the AI participants in the AI lifecycle. d) Determining ethical and legal requirements for FAITH outcomes based on current and upcoming regulatory instruments, assessing their applicability to FAITH, identifying gaps, and making recommendations. e) Proposing and measuring human attributes that determine the maturity of the teams for handling trustworthiness challenges, threats and incidents f) Developing and documenting metrics and scales for measuring risks from various dimensions of AI trustworthiness where possible and applicable. g) Estimating AI trustworthiness risks considering technical, social, and human threats. h) Selecting measurements/controls for managing trustworthiness risks that are technical, social, behavioural, legal, and policy related.

The proposed FAITH AI_TAF and measurements in D.2.1 will be evaluated through workshops with domain experts, users, AI participants from pilot domains, and affected communities, and finalized in D2.2. Deliverables D2.1 and D2.2 will complete the work in T.2.1.

1.3 Key Considerations

As we delve into the intricacies of this project, it is essential to consider several key questions that will guide our exploration and implementation. Firstly, we need to identify the LSPs AI systems that will be used in each pilot and determine which dimensions of trustworthiness are crucial for our AI systems, particularly in the sectors utilizing these technologies. Furthermore, understanding the roles of the FAITH AI participants—including AI participants in the Large Scale Pilots (LSPs) and affected communities—is vital; we must clarify their contributions and the specific phases of the AI lifecycle in which they will engage. In addition, it is important to outline the relevant legislation, standards, and policies that govern the sectors involved in the pilots. We must also identify key considerations for ensuring trustworthy and human-acceptable AI in the LSPs, including any potentially conflicting requirements and prioritizing associated threats and consequences. Finally, we should strategize on how to effectively attract FAITH AI participants to foster engagement and collaboration throughout the project.

2 Trustworthiness of AI systems and AI participants: Concepts

In this chapter, the basic concepts are analysed, where a glossary of terms (in Appendix C) accompanies it.

2.1 Analysis of trustworthiness in AI

The concept of "trustworthy AI" has its origins rooted in the broader discourse surrounding the recent success of AI technologies. As AI systems began to demonstrate significant capabilities and potential consequences across various sectors, an urgent need emerged to ensure that these systems operated in a manner that was reliable, fair, and aligned with societal values.

A pivotal moment for the formalization of trustworthy AI was the European Commission's launch of its AI strategy in 2018. This strategy marked the first comprehensive effort to articulate the principles and frameworks necessary to guide AI development within a regulatory context. The European Commission emphasized the dual pillars of **excellence** and **trust** as the foundation for Europe’s approach to AI. The goal was to harness the benefits of AI while addressing and mitigating potential risks, which included concerns about safety, ethical considerations, and the protection of fundamental human rights.

The strategy called for creating robust regulatory frameworks aimed at fostering innovation while ensuring that AI technologies were transparent, accountable, and aligned with public interest. Key components of this approach included consultations with a broad spectrum of stakeholders, including industry leaders, academia, civil society, and policymakers, to develop guidelines that would underpin trustworthy AI.

One of the significant outcomes of these efforts was the establishment of the High-Level Expert Group on AI, which published guidelines highlighting the essential requirements for trustworthy AI. These requirements encompassed principles such as transparency, accountability, fairness, and robustness, all crucial for earning and maintaining public trust in AI systems.

Additionally, the European Union took concrete legislative steps towards trustworthy AI with the AI Act, the first of its kind legal framework aimed at regulating AI technologies. This legislation seeks to ensure that AI applications are developed and deployed in ways that are safe and respectful of fundamental rights, thereby paving the way for AI systems that European citizens can trust.

Globally, the notion of trustworthy AI has been echoed in various initiatives, including frameworks and guidelines from other international organizations and national bodies. For instance, the U.S. National Institute of Standards and Technology (NIST) has put forth an AI Risk Management Framework, while the G7 Leaders have agreed upon international guiding principles as part of the Hiroshima Process on Artificial Intelligence. These efforts collectively contribute to the establishment of global standards for trustworthy AI, emphasizing the critical importance of ethics, transparency, and accountability in AI development.

Among the numerous initiatives and proposals emerged at national, regional and international levels to suggest possible ways and options for regulating and standardising the development of AI systems, we selected some key ones and report a summary below.

HLEG Guidelines on Trustworthy AI

The High-Level Expert Group on Artificial Intelligence (HLEG) established by the European Commission outlines principles for trustworthy AI, which include human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, environmental and societal well-being, and accountability. These guidelines serve as a framework to ensure AI systems are aligned with ethical standards and legal norms.

EU AI Act

The European Union's AI Act is a legislative framework aimed at regulating AI to ensure it is safe and respects fundamental rights. It introduces a risk-based approach categorizing AI applications into different levels of risk and imposing varying levels of regulatory control. High-risk AI systems will face strict requirements including transparency, robustness, accuracy, and cybersecurity, along with human oversight.

White House Blueprint for an AI Bill of Rights

The White House's AI Bill of Rights provides a set of five principles to guide the design, use, and deployment of AI systems. These principles include the right to be protected from unsafe or ineffective systems, the right to avoid discrimination, the right to privacy, the right to notice and explanation, and the right to human alternatives and fallback mechanisms, ensuring AI technologies align with democratic values and human rights.

Canada Bill for AI Regulation

Canada's AI regulation initiative focuses on promoting innovation while ensuring that AI systems are developed and used in ways that uphold human rights, inclusivity, transparency, and accountability. It includes provisions for mandatory risk assessments, consequence assessments, and adherence to ethical guidelines to prevent harm and ensure the fair treatment of all individuals.

ENISA AI Cybersecurity Challenges

The European Union Agency for Cybersecurity (ENISA) explores the cybersecurity challenges associated with AI. This includes ensuring the integrity, robustness, and resilience of AI systems, protecting against adversarial attacks, and establishing secure development and deployment practices. ENISA provides guidelines to mitigate risks and enhance the security of AI technologies.

OECD AI Recommendation

The Organisation for Economic Co-operation and Development (OECD) AI Principles advocate for AI systems that are innovative, trustworthy, and respect human rights and democratic values. The guidelines focus on inclusive growth, sustainable development, human-centered values and fairness, transparency and explainability, robustness, security and safety, and accountability.

Responsible AI Certification

Responsible AI Certification programs aim to establish standards and benchmarks for evaluating AI systems. These certifications ensure that AI solutions adhere to ethical guidelines and best practices, covering aspects such as fairness, transparency, accountability, and privacy, thus fostering trust and reliability in AI technologies.

White Paper on Trustworthy Artificial Intelligence by the China Academy for Information and Communication Technology (CAICT)

CAICT's white paper on Trustworthy AI discusses the principles and strategies for developing AI that is reliable, safe, and aligned with societal values. It emphasizes governance frameworks, ethical standards, and technical guidelines to ensure AI systems operate transparently and fairly, aligning with China's broader approach to AI regulation.

Deloitte Trustworthy AI Process

Deloitte's Trustworthy AI framework is a comprehensive approach to evaluate and ensure the reliability of AI systems. It includes ethical standards and governance protocols focused on fairness, transparency, accountability, reliability, and security. Deloitte provides strategic guidance for companies to develop AI technologies that meet these standards.

Conformity Assessment for Trustworthy AI

Conformity assessment involves evaluating AI systems to ensure they comply with regulatory and ethical standards. This process includes auditing AI development processes, verifying compliance with industry standards, and certifying that AI systems meet predefined criteria for safety, fairness, and transparency.

CEN-CENELEC and ISO/IEC

CEN (European Committee for Standardization) and CLC (European Committee for Electrotechnical Standardization), in collaboration with ISO (International Organization for Standardization) and IEC (International Electrotechnical Commission), have been working on various standardization initiatives for artificial intelligence (AI). These efforts are primarily aimed at ensuring the safe, ethical, and effective use of AI technologies across different sectors.

Key initiatives typically include:

1. **Ethical Guidelines**: Establishing clear principles for the ethical development and deployment of AI, such as transparency, fairness, accountability, and privacy considerations.
2. **Interoperability**: Creating standards that ensure AI systems and technologies can operate seamlessly across different platforms and environments.

3. **Safety and Security**: Developing safety standards to mitigate risks associated with AI, including cybersecurity threats and potential biases in AI algorithms.
4. **Performance Metrics**: Defining performance criteria to evaluate the effectiveness and reliability of AI applications.
5. **Terminology and Concepts**: Standardizing the terminology and basic concepts related to AI to facilitate a common understanding among stakeholders, including developers, regulators, and users
6. **Implementation Guidelines**: Providing frameworks and best practices for the practical implementation of AI systems in various industries.

CEN/CLC/JTC standardization initiatives for artificial intelligence. These initiatives are focused on developing European standards for AI, often in alignment with international standards.

1. CEN/CLC/JTC 21 - Artificial Intelligence: This joint technical committee was established to develop and provide standards for AI and related data. Its main objectives include:

   a) Developing standards for AI systems in various domains

   b) Addressing ethical concerns and societal consequence of AI

   c) Ensuring AI trustworthiness, including safety, security, and privacy

   d) Aligning with international standards (ISO/IEC) where appropriate

2. AI Ethics Guidelines: CEN and CENELEC are working on translating the EU's Ethics Guidelines for Trustworthy AI into concrete standards. This includes:

   a) Developing guidelines for implementing ethical AI principles

   b) Creating assessment lists for trustworthy AI

3. AI Risk Management Framework: Developing standards to help organizations identify, assess, and mitigate risks associated with AI systems.
4. AI Transparency and Explainability: Creating standards to ensure AI systems are transparent and their decisions can be explained to users and stakeholders.
5. AI Data Quality: Developing standards for ensuring the quality and integrity of data used in AI systems.
6. AI Governance: Working on standards for organizational governance of AI systems, including accountability and responsibility frameworks.
7. AI in Critical Infrastructure: Developing standards for the use of AI in critical infrastructure and high-risk applications.

8. AI Conformity Assessment: Creating frameworks for assessing AI systems' conformity to relevant standards and regulations.
9. AI Terminology and Concepts: Establishing a common vocabulary and conceptual framework for AI to ensure clear communication across the European AI landscape.
10. AI Performance Metrics: Developing standardized metrics for evaluating the performance of AI systems across different applications.

These initiatives aim to create a harmonized approach to AI standardization across Europe, supporting the EU's goal of becoming a global leader in trustworthy AI. They are designed to complement and sometimes localize international standards while addressing specific European needs and values.

1. ISO/IEC TR 24028: This technical report provides an overview of trustworthiness in artificial intelligence systems. It covers various aspects such as robustness, resiliency, reliability, accuracy, safety, security, and privacy.
2. ISO/IEC TR 24027: Focuses on bias in AI systems and AI-aided decision making. It provides guidance on identifying and addressing bias throughout the AI lifecycle.
3. ISO/IEC TR 24372: This report offers an overview of computational approaches for AI systems, including machine learning, reasoning, and knowledge representation.
4. ISO/IEC TR 24030: Provides use cases and applications of AI across various industries and domains.
5. ISO/IEC 22989: Defines key concepts and terminology related to artificial intelligence to establish a common language for AI discussions and development.
6. ISO/IEC 23053: Addresses the framework for artificial intelligence systems using machine learning, providing guidance on their development and implementation.
7. ISO/IEC TR 24368: Deals with ethical and societal concerns related to AI, offering guidance on addressing these issues in AI development and deployment.
8. ISO/IEC AWI TR 5469: Focuses on functional safety and AI systems, providing guidance on integrating AI into safety-critical applications.
9. ISO/IEC CD 42001: Aims to establish a governance framework for the development and use of artificial intelligence.
10. ISO/IEC WD 38507: Addresses governance implications of the use of AI by organizations, offering guidance on managing AI-related risks and opportunities

NIST Risk Management Framework for Artificial Intelligence

The National Institute of Standards and Technology (NIST) Risk Management Framework for AI provides guidelines for identifying, assessing, and managing risks associated with AI systems. It encourages practices that enhance the safety, fairness, and reliability of AI, incorporating insights into risk mitigation and promoting secure AI development processes.

FUTURE-AI Guiding Principles

The FUTURE-AI guiding principles emphasize a proactive approach to developing AI that is Fair, Universal, Transparent and Traceable, Usable, Robust, and Explained. These principles advocate for the adoption of practices and policies that ensure AI development is aligned with societal needs and ethical standards, fostering trustworthy and accountable AI systems.

Trustworthiness Dimensions-Aligning Definitions

Most of the above-mentioned policy initiatives and guidelines put forth common elements as Trustworthy AI dimensions. In the following tables, a summary of the main contributions and dimensions proposed:

Table 1: Trustworthy AI dimensions

	Technical Design Characteristics	Socio-Technical Characteristics	Guiding Principles Contributing to Trustworthiness
NIST AI RMF Taxonomy	Accuracy Reliability Robustness Resilience or ML Security	Explainability Interpretability Privacy Safety Managing Bias	Fairness Accountability Transparency
OECD	Robustness Security	Safety Explainability	Traceability to human values Transparency and responsible disclosure Accountability
EU AI Act EC HLEG	Technical robustness	Safety Privacy Non-discrimination	Human agency and oversight Data governance Transparency Diversity and fairness Environmental & societal well-being Accountability
EO 13960	Purposeful and performance-driven Accuracy, reliability and effectiveness Security and resilience	Safety Understandability by subject matter experts, users, and others, as appropriate	Lawfulness and respect of our Nation's values Responsibility and traceability Regular monitoring of performance Transparency Accountability
FUTURE-AI	data protection, AI risk management, AI evaluation planning, socio-ethical and legal awareness	Stakeholders' engagement, Explainability interpretability, usability, privacy-preserving	Fairness, Universality, Transparency, Usability, Robustness, Explainability

The summary of key dimensions that we draw from them are included in the following table.

Table 2: Key dimensions

Term	Pillar	Initiative
Accountability	Guiding Principles Contributing to Trustworthiness	AI RMF, OECD, EU AI Act, E 13960
Fairness & Non-discrimination	Guiding Principles Contributing to Trustworthiness	AI RMF, EU AI Act, FUTURE-AI
Transparency	Guiding Principles Contributing to Trustworthiness	AI RMF, OECD, EU AI Act, FUTURE-AI, EO 13960
Privacy & Data Governance	Privacy & Data Governance	AI RMF, EU AI Act
Explainability & Interpretability	Socio-Technical Characteristics	AI RMF, OECD, EO 13960
Safety & Security	Socio-Technical & Technical Design Characteristics	AI RMF, OECD, EU AI Act, FUTURE-AI, EO 13960
Reliability & Robustness	Technical Design Characteristics	AI RMF, OECD, EU AI Act, FUTURE-AI
Traceability & Auditability	Guiding Principles Contributing to Trustworthiness	AI RMF, OECD, FUTURE-AI, EO 13960
Human Agency & Oversight	Guiding Principles Contributing to Trustworthiness	EU AI Act
Environmental & Societal Well-being	Guiding Principles Contributing to Trustworthiness	EU AI Act

These characteristics collectively contribute to building trustworthy AI systems that can be deployed ethically and effectively across various applications and industries.

2.1.1 Dimensions of trustworthiness-AI Participants

In the fast-changing realm of artificial intelligence (AI), establishing trust is crucial for broad acceptance and successful integration across diverse fields. The National Institute of

Standards and Technology (NIST) AI Risk Management Framework (AI RFM) outlines key attributes that determine the reliability of AI systems:

Fit for Purpose: The idea of being "fit for purpose" emphasizes the necessity of aligning the design and functionalities of an AI system with its intended goals. This aspect ensures that the system not only demonstrates technical competence but also effectively meets the requirements of its users. Within the realm of AI ethics and design, concepts such as fairness, accountability, and transparency are pivotal in creating a system that genuinely fulfils its intended purpose [3].

Predictable and Dependable: Consistency in AI behaviour is a crucial attribute that allows users to anticipate how the system will respond and perform in various circumstances. To achieve this consistency, transparency in AI algorithms and decision-making processes is essential, as it helps users understand and have confidence in how the system operates [4]. Dependability, meanwhile, refers to the system's ability to maintain consistent and reliable performance over time, thereby minimizing the chances of unexpected errors or deviations from established norms [5].

Appropriate Level of Automation: Maintaining an appropriate level of automation is crucial for ethically and reliably integrating AI. This aspect recognizes the boundaries of AI systems and underscores the significance of human supervision, especially in intricate or morally nuanced scenarios. Finding this equilibrium guarantees that AI enhances human capabilities while retaining oversight, thereby fostering conscientious and accountable AI implementations [6].

The dimensions of AI trustworthiness outlined in the NIST AI RFM include suitability for purpose, predictability, reliability, and maintaining an optimal level of automation. These dimensions establish a comprehensive framework to steer the design and implementation of AI systems, ensuring alignment with ethical standards and meeting the needs of users and stakeholders across various domains.

2.1.2 AI threats and risks versus traditional ICT ones

AI threats differ significantly from traditional software threats due to the complexity and autonomy inherent in AI systems. Unlike conventional software, AI systems are dynamic systems that can exhibit unpredictable behaviours, propagate biases, and amplify errors on a larger scale, impacting not only organizational operations but also societal dynamics. The reliance on vast datasets and sophisticated algorithms introduces unique challenges in terms of accountability, transparency, and ethical considerations. Addressing AI risks requires a nuanced understanding of these factors to develop robust mitigation strategies that safeguard against potential harm and ensure responsible deployment and use of AI technologies.

Like conventional software, risks associated with AI technology can extend beyond individual enterprises, affecting multiple organizations and even broader societal realms. AI systems introduce risks that current frameworks and methods do not fully address. However, certain features of AI systems, such as pre-trained models and transfer learning, offer potential benefits by enhancing research capabilities and improving accuracy and robustness compared to traditional models. Understanding contextual factors within the MAP (Maximum a Posteriori) function, which is a probabilistic approach used for estimating AI model parameters, will help AI participants assess the level of risk and devise effective management strategies.

AI-specific risks that are either new or heightened include:

- When constructing AI systems, the data used may not effectively capture their intended context or purpose, and establishing a clear ground truth can be challenging. This lack of accuracy, compounded by issues such as biased data and quality problems, undermines the reliability of AI systems. Consequently, there is a risk of adverse consequences stemming from these reliability issues.
- AI systems' reliance on training data, which is often extensive and complex, increases their vulnerability to variations in data quality.
- The performance of AI systems can be significantly affected by alterations that occur during training, whether they are deliberate or inadvertent.
- Over time, the datasets utilized to train AI systems may diverge from their initial context or become obsolete compared to the scenarios in which they are deployed.
- AI systems frequently exhibit complexity, integrating billions or even trillions of decision points within conventional software frameworks.
- The use of pre-trained models, which enhances research and performance, also introduces challenges in handling statistical uncertainty, bias, scientific validity, and reproducibility.
- Predicting failure modes and emergent properties in large-scale pre-trained models presents significant challenges.
- The improved data aggregation capabilities of AI systems increase concerns regarding privacy risks.
- The need for maintenance of AI systems may grow as a result of data, model, or concept drift over time.
- Opacity in AI systems raises concerns regarding their reproducibility and transparency.
- The standards for testing software in AI-based practices are still in development, lacking the comprehensive documentation typically found in traditionally engineered software.
- Regularly testing AI-based software poses challenges due to the distinct control mechanisms involved, which differ significantly from those in traditional code development.
- The costs involved in developing AI systems computationally can lead to substantial environmental consequences.
- It is difficult to anticipate or detect the unintended consequences of AI systems that go beyond statistical measurements.
- The human element threats have not been identified.

AI risks present a stark contrast to traditional Information and Communication Technology (ICT) risks due to the distinctive features and capabilities of AI systems. Unlike conventional ICT systems, AI systems have the capacity for autonomous learning and adaptation, which makes their behaviours less predictable and their potential consequences more profound. Traditional ICT risks often revolve around issues such as data breaches, system failures, and network vulnerabilities, which are typically mitigated through established cybersecurity measures and protocols. In contrast, AI risks encompass a broader spectrum, including ethical concerns around bias and fairness, the opaque nature of decision-making processes, and the potential socio-economic consequences of widespread automation. Addressing AI risks requires a multifaceted approach that combines technical expertise with ethical considerations and regulatory frameworks tailored to the unique challenges posed by artificial intelligence. As AI continues to evolve and integrate deeper into societal and economic frameworks, understanding and mitigating these risks will be crucial in harnessing the full potential of AI technology while safeguarding against its unintended consequences.

2.1.3 AI Lifecycle

The AI lifecycle (Figure 1) follows a cyclical process involving three main stages: Design, Testing (Develop), and Production (Deploy). In the Design phase, the primary focus is on understanding the problem that needs to be addressed, gathering and exploring relevant data, and preparing the data through cleansing and normalization procedures. This phase sets the foundation for the subsequent steps by ensuring a clear definition of the problem and a robust dataset. The Testing phase involves modelling and evaluation, where various machine learning algorithms are applied to the prepared data to create models. These models are then evaluated to ensure their accuracy and effectiveness. Finally, in the Production phase, the validated models are moved into production where they are monitored to ensure they perform as expected in real-world scenarios. This phase includes continuous monitoring and maintenance to address any issues that arise and to ensure the model's performance remains optimal. The cyclical nature of this lifecycle highlights the iterative process of refining and improving AI models based on new data and insights.

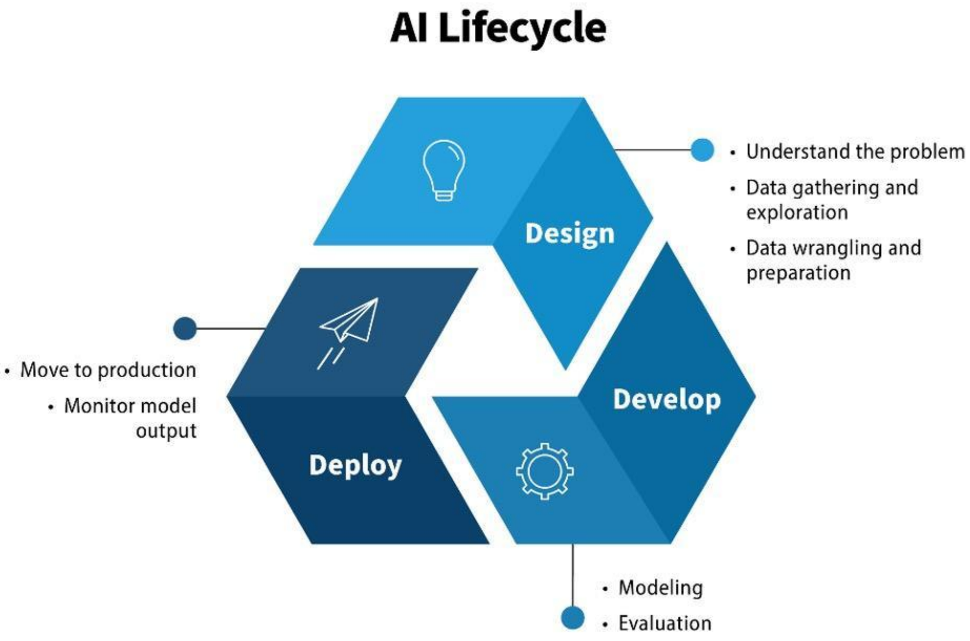


Figure 1: Visual overview of the AI lifecycle.

2.2 Human element in the AI trustworthiness

There are two types of humans that play a crucial role in the AI trustworthiness: the organizations' teams with AI participants (within the AI life cycle) and the potential adversaries.

2.2.1 AI participants in the organization's teams within the AI lifecycle

Furthermore, AI participants are the key players involved in the creation, design, development, and implementation of artificial intelligence systems. They encompass a range of roles including designers, developers, and data specialists, all working together to ensure that AI technologies are effective, ethical, and aligned with their intended purposes. The AI participants related to the FAITH AI_TAF are the following [2] (Figure 2):

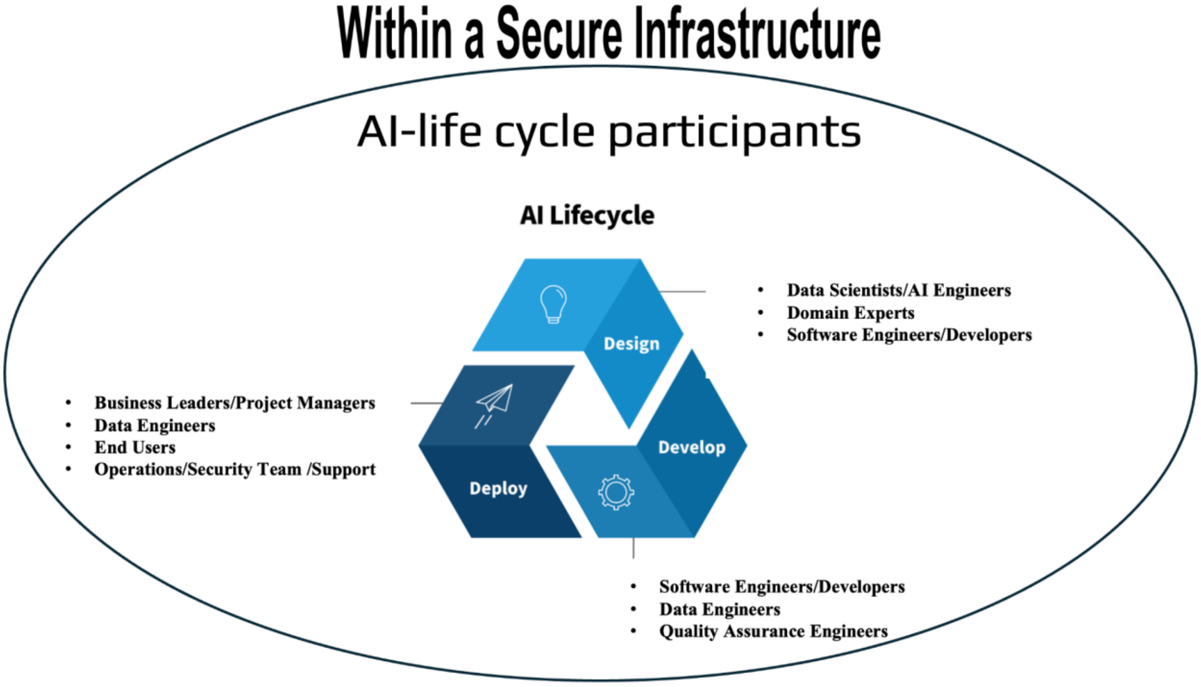

Within a Secure Infrastructure
AI-life cycle participants
AI Lifecycle
Design
Develop
Deploy
Data Scientists/AI Engineers
Domain Experts
Software Engineers/Developers
Business Leaders/Project Managers
Data Engineers
End Users
Operations/Security Team /Support
Software Engineers/Developers
Data Engineers
Quality Assurance Engineers

Figure 2: AI participants (ENISA, 2023).

AI designers are responsible for conceptualizing and setting the goals for AI systems. They handle the planning, designing, and data-related tasks to ensure that the AI systems they create are both compliant with legal standards and tailored to their intended functions. This category of AI participants encompasses a diverse group including data scientists, domain specialists, socio-cultural analysts, diversity and inclusion experts, representatives from affected communities, human factors professionals, governance authorities, data engineers, data suppliers, funding entities, product managers, external organizations, evaluators, and legal and privacy advisors.

AI Development participants establish the foundational infrastructure for AI systems and are tasked with building and interpreting models. This includes creating, selecting, calibrating, training, and testing models or algorithms. Key players in this category include machine learning specialists, data scientists, developers, external organizations, legal and privacy governance professionals, and experts in socio-cultural and contextual aspects relevant to the deployment environment.

AI Deployment participants handle the contextual decisions regarding the utilization of AI systems to ensure their successful implementation into production. Their responsibilities include piloting the system, ensuring compatibility with existing systems, maintaining regulatory compliance, managing organizational changes, and assessing user experience. This group includes system integrators, software developers, end users, operators and practitioners, evaluators, and domain experts in human factors, socio-cultural analysis, and governance.

AI Operation and Monitoring participants are linked with the operation of AI systems and collaborate with others to continuously evaluate the system's outputs and its broader consequences. In this category, AI participants include system operators, domain experts, AI designers, users who interpret or integrate AI system outputs, product developers, evaluators and auditors, compliance specialists, organizational management, and researchers within the community.

AI Test participants specialize in examining the AI system or its components, identifying issues, and implementing solutions to ensure smooth operation.

AI Human Centred participants. Human factors professionals contribute diverse skills and perspectives to comprehend the context of use, promote interdisciplinary and demographic diversity, engage in consultative processes, design and assess user experiences, conduct human-centred evaluations and tests, and inform consequence assessments.

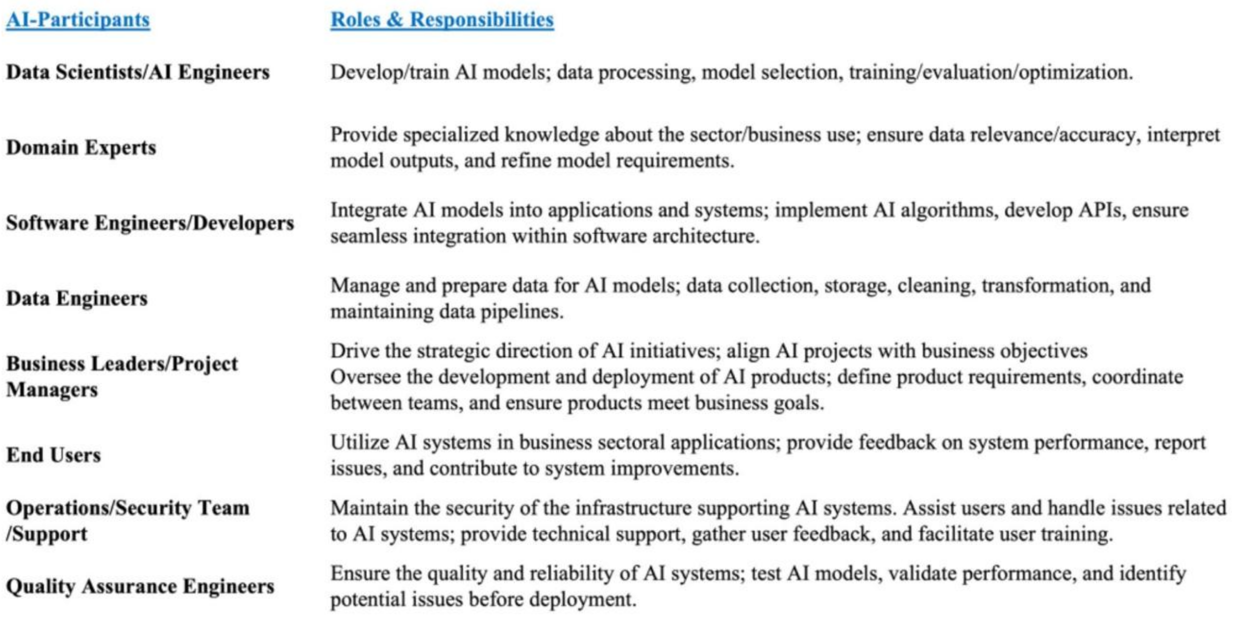
AI Domain expert participants. Domain experts among AI participants play a crucial role in guiding the design and development of AI systems. They also interpret outputs to support the efforts of Testing, Evaluation, Validation, and Verification teams and AI consequence assessment teams.

AI Consequence Assessment involves evaluating various aspects such as AI system accountability, addressing biases, assessing consequences on product safety, liability, and security. AI participants specializing in consequence assessment and evaluation contribute technical, human factors, socio-cultural, and legal expertise to these tasks.

Third-party entities encompass providers, developers, vendors, and evaluators who offer data, algorithms, models, and/or systems, along with associated services to organizations or their clients. These entities assume responsibility for AI design and development tasks, either wholly or partially.

End users of an AI system are individuals or groups who utilize the system for particular purposes within a specific context. They vary widely in expertise, ranging from AI specialists to those encountering technology for the first time.

Other AI participants may establish formal or semi-formal standards or guidelines for defining and mitigating AI risks (Figure 3). These participants encompass trade associations, standards development organizations, advocacy groups, researchers, environmental organizations, and civil society organizations.



AI-Participants	Roles & Responsibilities
Data Scientists/AI Engineers	Develop/train AI models; data processing, model selection, training/evaluation/optimization.
Domain Experts	Provide specialized knowledge about the sector/business use; ensure data relevance/accuracy, interpret model outputs, and refine model requirements.
Software Engineers/Developers	Integrate AI models into applications and systems; implement AI algorithms, develop APIs, ensure seamless integration within software architecture.
Data Engineers	Manage and prepare data for AI models; data collection, storage, cleaning, transformation, and maintaining data pipelines.
Business Leaders/Project Managers	Drive the strategic direction of AI initiatives; align AI projects with business objectives Oversee the development and deployment of AI products; define product requirements, coordinate between teams, and ensure products meet business goals.
End Users	Utilize AI systems in business sectoral applications; provide feedback on system performance, report issues, and contribute to system improvements.
Operations/Security Team /Support	Maintain the security of the infrastructure supporting AI systems. Assist users and handle issues related to AI systems; provide technical support, gather user feedback, and facilitate user training.
Quality Assurance Engineers	Ensure the quality and reliability of AI systems; test AI models, validate performance, and identify potential issues before deployment.

Figure 3: AI participants for defining and mitigating AI risks.

2.2.2 Personality Attributes of AI participants and Organizational Maturity

Trustworthy AI requires a sufficient level of organizational maturity, particularly in the teams of AI participants. Such maturity may depend on a range of factors, including organizational processes, culture, and individual training. Scientific literature shows that there may also be a link between individual human personality traits (See table 3 below) and the trustworthiness towards IT and AI systems. The ability of an organisation to manage AI related challenges depend upon the maturity of its teams in cultivating trustworthy AI.

Table 3: Collective Personality Attributes and Traits of AI teams

Category	Description
Personality Traits	- Vigilance: Alert and attentive to AI threats. - Responsibility and curiosity takes ownership, driven by curiosity. - Adaptability: open to new technologies and strategies. - Openness to Experience: intellectual, creative and adventurous - Resilience: copes well with stress and setbacks
Social Traits	- Conventional relationships: adapts to social norms, builds strong bonds - Collaboration: works effectively with teams and partners. - Professional virtual relationships establishes virtual collaborations easily. - Ethics: prioritise honesty, transparency and respect.
Soft Skills	- Problem solving and teamwork: strong analytical and communication skills.

	- Cross functional collaboration works with diverse teams. - Documentation: proficient in creating concise security documentation. - Continuous learning: new trends, tools, practices.
Available Resources	- High performance computing, big databases, modelling tools etc. - Community involvement in AI: active in AI security communities (CERT, ISACS) and AI standardisation bodies (ETSI, CEN, etc.)
Relationships with the AI lifecycle	Data scientists/ AI engineers, Domain experts, Software engineers/ Developers, Data engineers, Business Leaders/ Project Managers, End Users, Operations/ Security Teams (Risk Assessors)/ Support Quality Assurance engineers.
Motivations for Trustworthy Behaviour	Protecting AI systems: adheres to ethics, respects privacy, security, transparency, non-bias and legal standards. - Fostering trust follows trustworthy AI best practices. - Public safety, Organisational security, trustworthy AI services and products enhances security attitude, withstand to cybercrime. - Continuous advancement: focuses on skill and knowledge improvement.
Cognitive skills	- Cognitive flexibility: adapts to new information and challenges. - Creativity: generates innovative solution. - Information processing: analyses data, retains and recalls information.

2.2.3 AI adversaries

AI stakeholders must understand their adversaries within the operational context. Potential adversaries are defined by three essential elements: **means, motive, and opportunity.** An attack takes place when the adversary possesses the means to carry it out, seizes an opportunity to exploit vulnerabilities, and harbours the motive to target the specific victim involved.

AI stakeholders and operators must assess potential attackers to more accurately gauge their risk levels and implement suitable countermeasures [7].

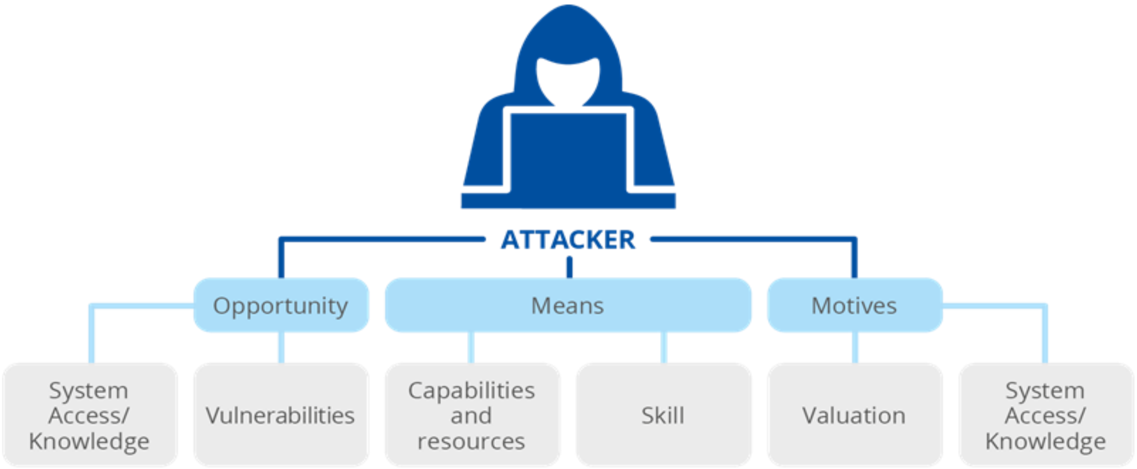


Figure 4: Adversary characterization (ENISA, 2023).

Adversarial threats [8] primarily stem from individuals who intentionally seek to inflict harm. These individuals are commonly known as attackers or adversaries. In literature, efforts are ongoing to develop cyber threat actor lists and taxonomies, which typically categorize intentional threat actors into several groups: insider attackers, cyber terrorists, hacktivists/civil activists, organized cybercriminals, script kiddies, state-sponsored attackers, commercial industrial espionage agents, cyber warriors/individual cyber fighters, cyber vandals, and black hat hackers.

Currently, there is no globally recognized standard for an attacker taxonomy, and new definitions and proposals for taxonomies continue to emerge. ENISA [8] defined 11 types of attackers in 2021-2022, integrating and enhancing previous taxonomies to align with the evolving threat landscape. These classifications can be cross-referenced with taxonomies used by Member States and EU bodies. Attackers target ICT infrastructures that host AI systems/products or AI systems at any stage of their lifecycle.

Historically, assessment methods, such as those in ISO/IEC and NIST frameworks, have focused primarily on the technical skills, resources, and motivations of attackers. While these assessments provide some insights, they often overlook the human factors that can influence an attacker's behaviour. Integrating a broader range of traits and competencies into risk assessments can offer a more comprehensive evaluation of potential threats.

trustilio FAITH partner extended existing methodologies by developing an "attackers' profile," which is multidimensional and based on factors from various scientific fields, including psychology, criminology, and cyberpsychology. This profile incorporates key personality traits from the Five-Factor Theory (Table 4), such as agreeableness, extraversion, conscientiousness, neuroticism, and openness to experience, as well as behavioural tendencies described by Fogg's Behaviour Model (B=MAT, Behaviour = Motivation + Ability + Trigger). These models are used to assess an attacker's likelihood of executing a cyberattack by analysing not only their technical capabilities but also their psychological, social, and motivational traits.

Table 4: Facets of the Five-Factor Theory (FFT) Model

Traits	Facet	Example
Agreeableness	Trust	Trusting other people in collaboration
Extraversion	Positive Emotions	High energy in social engagements
Conscientiousness	Self-efficacy	Strong focus on achieving goals
Neuroticism	Self-consciousness	Awareness of social behaviours
Openness to Experiences	Creativity	Innovation and imaginative thinking

In addition to these psychological factors, attackers' social and behavioural traits, such as ease in forming anonymous relationships in hacking communities or the ability to manipulate others (e.g., through phishing), technical traits, such as networking and IT skills, and the ability to exploit vulnerabilities, are also crucial in understanding an attacker's overall profile (see Table 5 and extended list in section 2.2.4 below).

Table 5: Social, Behavioural, and Technical Traits of Attackers

Category	Traits	Description & Example
Social	Anonymous relationships	Forming virtual professional bonds on the Deep Web
Behavioural	Manipulation	Manipulating people via electronic means (phishing)
Technical	Networking skills	Knowledge in DNS, HCP, and systems architecture
Motivational	Personal Satisfaction	Motivated by personal goals (competition, boredom)

Based on this taxonomy, trustilio proposed a scoring system for attackers' profiles reflecting on the percentage of traits exhibited across psychological, social, and technical domains. This scoring system classifies attackers into categories ranging from "Insufficient" to "Sophisticated" based on the extent to which they exhibit the key traits identified. The scoring is also linked to the Attack Potential (AP), allowing cybersecurity professionals to more accurately assess the likelihood and severity of an attack (Table 6).

Table 6: Scoring the Extended Attackers' Profile

Attackers' Profile	Qualitative Values	Semi-Quantitative Values	Percentage of Traits Exhibited
Sophisticated	10	96-100	More than 96% of traits

Experienced	8	80-95	More than 80% of traits
Moderate	5	21-79	21-79% of traits
Basic	2	5-20	5-20% of traits
Insufficient	0	1-4	Less than 5% of traits

The scoring system is then used to estimate the AP by linking the attackers' profiles with the potential consequence of their actions. For instance, attackers with a "Sophisticated" profile (such as nation-state actors or cyber-terrorists) who have both technical expertise and significant resources are more likely to carry out a successful attack, leading to an "AP Beyond High" classification (Table 7).

Table 7: Scoring Attack Potential (AP)

AP Qualitative Value	AP Quantitative Value	Description
Beyond High	10	Sophisticated Profile (multi-sectoral expert)
High	8	Experienced Profile
Moderate	5	Moderate Profile
Basic	2	Basic Profile
Very Low	0	Insufficient Profile

As such we have a more detailed method for profiling attackers and assessing their threat level. This interdisciplinary approach, which integrates human factors with technical capabilities, allows for better forecasting of attack likelihood and helps in selecting appropriate security controls to mitigate risks. The proposed methodology can be applicable in sectors relevant to FAITH, where critical infrastructures are often targeted by highly skilled attackers and constituted the base for the proposed approach in Sections 2.3.5.1 and 2.3.5.2 below.

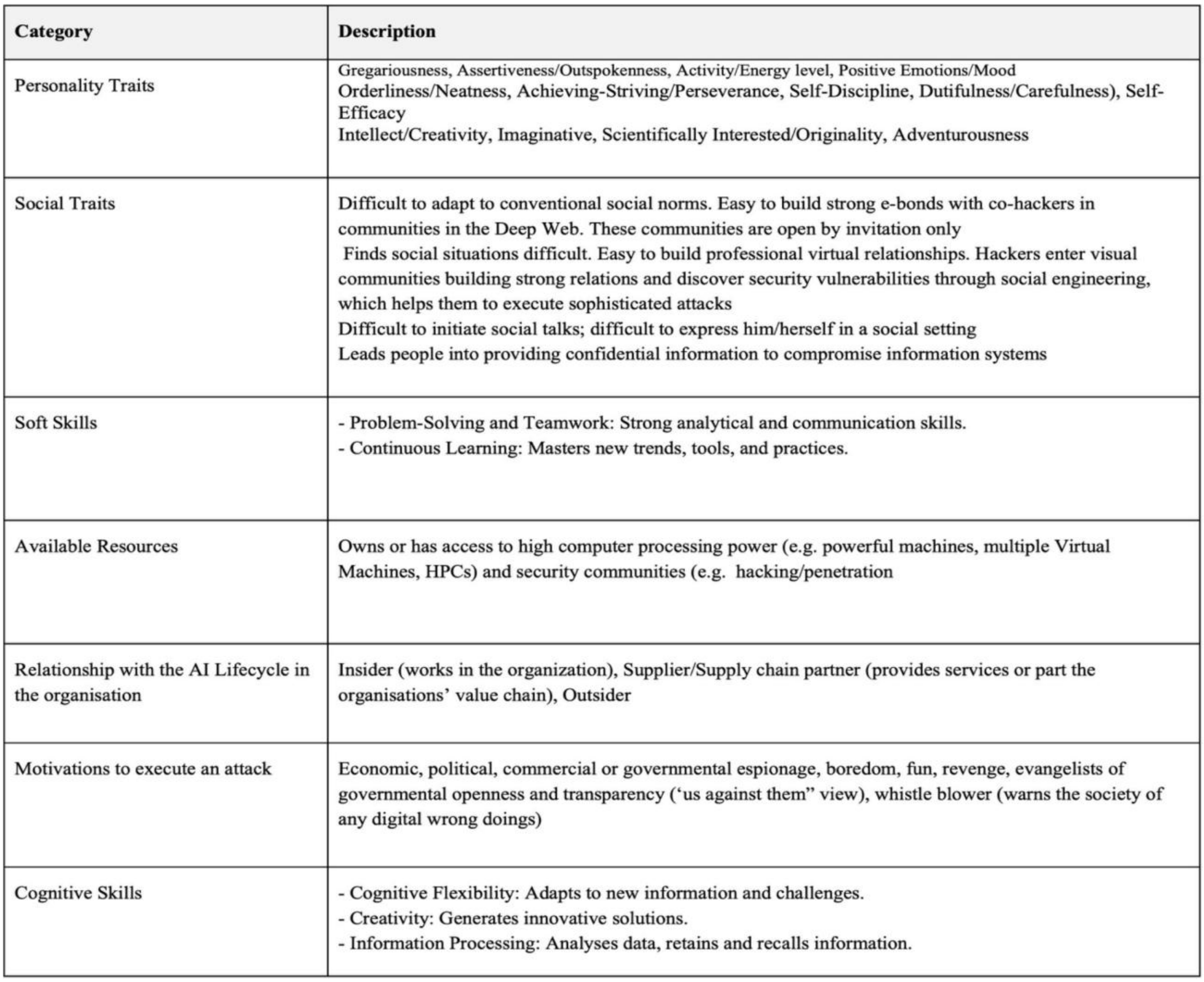
2.2.4 Profiles of AI adversaries (AP) and Traits

Understanding an AP requires a combination of research, intelligence gathering, and behavioural analysis. Attackers vary in type, including cybercriminals, terrorists, insider threats, hacktivists, and serial offenders, each with unique motivations and tactics.

To profile an attacker, one must define the type of attack, analyse methods and motives, and use Open-Source Intelligence (OSINT) tools such as Shodan, Maltego, and MITRE ATT&CK for digital tracking. Threat intelligence platforms and law enforcement reports (e.g., ENISA, EU-CERT, SOCs, Europol, Mandiant) provide deeper insights into attack patterns.

Psychologically, attackers often exhibit low empathy, manipulativeness, risk-taking, and high adaptability. Many possess Dark Triad traits (narcissism, Machiavellianism, psychopathy), which drive their malicious actions.

The following table outlines various categories of AI adversary traits.



Category	Description
Personality Traits	Gregariousness, Assertiveness/Outspokenness, Activity/Energy level, Positive Emotions/Mood Orderliness/Neatness, Achieving-Striving/Perseverance, Self-Discipline, Dutifulness/Carefulness), Self-Efficacy Intellect/Creativity, Imaginative, Scientifically Interested/Originality, Adventurousness
Social Traits	Difficult to adapt to conventional social norms. Easy to build strong e-bonds with co-hackers in communities in the Deep Web. These communities are open by invitation only Finds social situations difficult. Easy to build professional virtual relationships. Hackers enter visual communities building strong relations and discover security vulnerabilities through social engineering, which helps them to execute sophisticated attacks Difficult to initiate social talks; difficult to express him/herself in a social setting Leads people into providing confidential information to compromise information systems
Soft Skills	- Problem-Solving and Teamwork: Strong analytical and communication skills. - Continuous Learning: Masters new trends, tools, and practices.
Available Resources	Owns or has access to high computer processing power (e.g. powerful machines, multiple Virtual Machines, HPCs) and security communities (e.g. hacking/penetration
Relationship with the AI Lifecycle in the organisation	Insider (works in the organization), Supplier/Supply chain partner (provides services or part the organisations' value chain), Outsider
Motivations to execute an attack	Economic, political, commercial or governmental espionage, boredom, fun, revenge, evangelists of governmental openness and transparency ('us against them" view), whistle blower (warns the society of any digital wrong doings)
Cognitive Skills	- Cognitive Flexibility: Adapts to new information and challenges. - Creativity: Generates innovative solutions. - Information Processing: Analyses data, retains and recalls information.

Figure 5: Profiles and Traits of AI adversaries.

2.2.5 AI attacks

Adversarial threats in AI are traditionally categorized based on their targets, timing of attacks, attacker knowledge, and resulting consequences [9]. Targets may include physical components such as manipulated sensor inputs, digital representations like modified pre-processed data inputs, or the AI model itself, such as attacks on classification models. Attacks can occur during the training or inference stages. Attackers' knowledge ranges from extensive (white-box attacks) to minimal or none (black-box attacks), depending on their understanding of the model's architecture, parameters, training methods, and data. These attacks can compromise the integrity, availability, and confidentiality of AI systems. Figure 6 illustrates various attack types across different phases of the AI lifecycle.

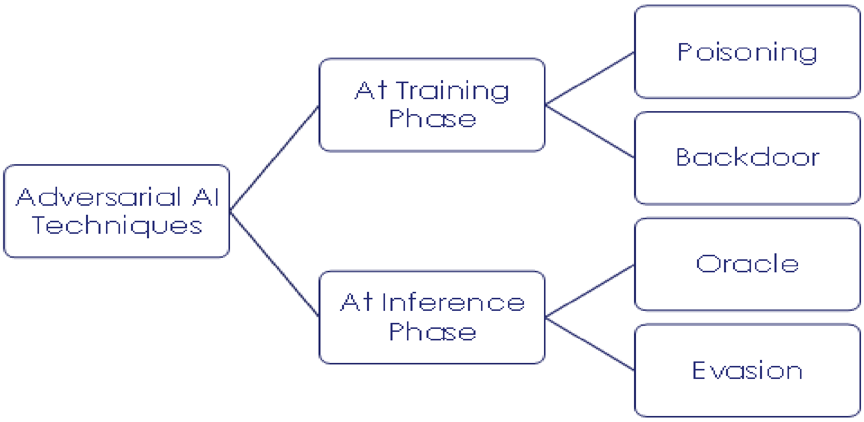


Figure 6: Attacks on AI.

Below, a summary of recognized adversarial AI methods based on these characteristics is presented.

Attacks such as data access, poisoning, and backdoors occur during the training phase. For instance, with access to training data (e.g., leaked or publicly available), an attacker could construct an alternative AI model to use for future attack strategies [9]. This type of attack necessitates knowledge of the training data, classifying it as either a white-box or grey-box attack, contingent on the attacker's familiarity with the model. Data access primarily consequences confidentiality, although it can lead to further consequences as a preliminary step to model attacks. Poisoning attacks involve methods to inject or manipulate training data [10]. Indirect poisoning occurs pre-processing and does not necessitate special privileges but does require a solid understanding of the application domain. Conversely, direct poisoning occurs post-processing and demands access to the training environment. Poisoning attacks generally affect the integrity and availability of AI models. Unlike data access or poisoning, AI backdoors do not affect the physical or digital representation but instead target the model itself. Side module insertion involves adding supplementary nodes to perform concealed tasks within a neural network architecture. Alternatively, deep alteration techniques introduce bias by modifying specific nodes [11]. Backdoors necessitate full knowledge and can consequence all system properties.

During the model inference stage, adversaries engage in evasion and oracle attacks. Evasion tactics involve adversaries seeking subtle alterations to inputs that lead to significant changes in output predictions [12]. Typically, gradient-based techniques manipulate computer vision systems to induce misclassification, while gradient-free methods provide alternatives if the AI employs gradient-masking techniques [13]. Evasion attacks may begin with limited initial knowledge, though black box attacks often require extensive trial and error. These attacks can target specific inputs to disrupt system performance or aim for broader consequences on integrity and availability [14]. In contrast, oracle attacks use model outputs and available data to deduce information about the model or its training data [15]. Membership inference checks if specific inputs were part of the training dataset [16], while inversion attempts to reconstruct training data and extraction aims to reverse-engineer the model. Oracle attacks allow adversaries to progress from black-box to gray and white-box knowledge levels [15], compromising the confidentiality of AI models and associated data.

3 A risk assessment-based approach to AI trustworthiness - State of the Art

This chapter presents an exploration of the state of the art in applying risk assessment methods to evaluate and optimize AI trustworthiness, particularly within the EU context. The discussion begins by examining key standards, frameworks, regulations, and policies, such as ISO CEN/CENELEC, the NIST AI RMF, and ENISA guidelines, to establish a foundation for trustworthiness assessment. The chapter further delves into current best practices and methodologies employed for AI trustworthiness evaluation.

In addition, it surveys relevant AI technologies and resources that contribute to enhancing trustworthiness. These include advanced AI/ML modelling techniques, decision intelligence approaches, serious games for testing and evaluation, anomaly detection methods, rules-based knowledge management systems, and approaches like anonymous human profiling. Through this comprehensive review, the chapter outlines the evolving landscape of AI trustworthiness and the risk-based strategies that are being integrated into various frameworks and practices.

3.1 AI risk assessment Frameworks

Risk assessment is key to understand, plan for, mitigate, and reduce the risks involved in AI systems. Different bodies and organizations have published and are developing AI risk management frameworks, best practices for, and standards regarding the development, deployment, and use of AI systems.

They include the “Multilayer Framework for Good Cybersecurity Practices for AI” by ENISA [17] and the “Artificial Intelligence Risk Management Framework” by NIST [18]. Whilst different AI risk management frameworks exist, they all share a similar basis and have commonalities between them.

3.1.1 ENISA’s Multilayer Framework for Good Cybersecurity Practices for AI

Amongst other aims, the ENISA framework “Multilayer Framework for Good Cybersecurity Practices for AI” aims for good AI cybersecurity practices that consider the whole AI lifecycle, and also to identify gaps in the already-existing AI cybersecurity practices. Within it, the need for AI-specific practices to compliment current cybersecurity practices is highlighted, as AI systems do not sit alone from cyber-physical systems and instead depend and run on them, and typically together they form part of a larger system of operation.

Data sources, data, algorithms, training models, implementation/data management/testing processes and users are identified as the main components to consider when treating AI systems as cyber assets within ICT infrastructure. Additionally, ENISA then uses a definition of an AI system from the OECD [19], that has since been updated [20] though remains compatible with the contextual use of it here and breaks down the types of AI systems into the multiple subfields of computer vision, expert systems, machine learning (ML), multi-agent systems, natural language processing, robotics, and speech recognition. Here the definition of an AI system used is 'a machine-based system that can influence the environment by producing an output (predictions, recommendations or decisions) for a given set of objectives. It uses machine and/or human-based data and inputs to: (i) perceive real and/or virtual environments; (ii) abstract these perceptions into models through analysis in an automated manner (e.g. with ML) or manually; and (iii) use model inference to formulate options for outcomes. AI systems are designed to operate with varying levels of autonomy.'

A three-layer scalable framework for good AI cybersecurity practices is proposed that categorises various identified best practices and standards such that they can be used by AI stakeholders to address the cybersecurity challenges of their AI systems.

First layer

The first layer, termed Cybersecurity Foundations, covers the good cybersecurity practices of the ICT environment that the AI system is hosted on.

Second layer

The second layer, AI Specific, covers the cybersecurity practices for the AI components in a sector-agnostic manner, considering component lifecycle, properties, threats, and controls. This also includes AI threat assessment and AI security management. The third layer accounts for the specific sector the AI system is used within.

ENISA gives the key elements of the second layer as the types of AI, AI assets and procedures, AI threat assessment, AI security management, AI-related standards, and trustworthy AI, amongst others. As ML has been at the forefront of the adoption of AI in different domains, and is used within a large amount of the different AI subcategories, a lifecycle is proposed that is based on ML. The different AI assets encompassing this are data, models, artefacts, actors/stakeholders, processes, and environment/tools.

AI risk assessments need to include both the traditional cyber-physical threats and threats that are specific to AI systems. For the case where EU jurisdiction applies, those AI-centric threats are mentioned in the EU AI Act: loss of transparency, interpretability, managing bias, and accountability must be included. Additionally, AI risk assessments must consider the robustness, resilience, fairness, and explainability of AI systems, and be dynamic such that they can be used in real-time with live AI systems and account for the anomalies that may be detected in them.

ETSI has an AI threat ontology [21] that defines what is an AI threat and how it differs from traditional cyber threats and ENISA categorise them.

Following the ENISA publication "Securing Machine Learning Algorithms" [22], ENISA also categorise the most important AI-based threats for ML systems as evasion, poisoning, model or data disclosure, compromise of ML application components, and failure or malfunction of an ML application. For an AI system, these threats can be mapped to multiple vulnerabilities and used within risk assessment.

An AI system will also have desired characteristics that, when present and not undermined, contribute to the user-perceived trustworthiness of an AI system. Defining AI trustworthiness as “the confidence that AI systems will behave within specified norms, as a function of some characteristics […]”, ENISA specifies different characteristics of AI trustworthiness.

Equivalent to security control of cyber-physical systems, security controls of AI systems can also be present and used effectively in the mitigation and prevention of AI-based threats, risks, and harms. For AI systems, related security controls minimise the compromise of the different AI characteristics and thus through their effective use an AI system is more likely to be considered trustworthy by its users and those who interact with it. ENISA discusses specific ML security controls that can be used to address the mentioned threat types of evasion, poisoning, and model or data disclosure. Additionally, ENISA highlights that AI security needs to be considered at all stages of the lifecycle, considering that AI systems are multi-disciplinary socio-technical systems, that ML and deep learning (DL) pose the main challenges, and that AI-specific risk assessment needs to consider the unique properties of AI systems. AI-relevant controls can be found in the ISO 2700x standards [23], NIST AI RMF, and ENISA’s best practices.

Third layer

For the third layer of the framework, ENISA considers sector-specific AI cybersecurity good practices. Here the energy, health, automotive, and telecommunications sectors are explored. They point out that threats exist in all economic sectors independent of how AI is being used, that the fragmented recommendations, best practices, solutions, and tools become stumbling blocks for sectoral stakeholders, and that collaboration and information sharing on sector-specific issues and mitigations between sectoral stakeholders is needed.

In their conclusions and way forward, ENISA highlights that for trustworthy AI systems, cybersecurity and AI experts need to continually assess the integrity of data sources and data, continually monitor the data lifecycle security, and perform dynamic AI lifecycle-encompassing risk assessment and management. They also point out that interdisciplinary expert collaboration is needed to develop trustworthy AI, and that globally accepted frameworks for AI ethics are needed.

Through this framework, a good basis for AI risk assessment and management is given, as are important concepts, definitions, and considerations.

3.1.2 NIST’s Artificial Intelligence Risk Management Framework

The goal of the NIST Artificial Intelligence Risk Management Framework is to be a resource to organizations designing, developing, deploying, or using AI systems to help them manage the risks of AI and to foster trustworthy AI. Risk management is key in the NIST framework to enhance the trustworthiness of AI and cultivate public trust.

The NIST AI RMF adapts the since-updated OECD 2019 definition [19] and the ISO definition [24] of an AI system to define an AI system as ‘an engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.'

NIST organises the AI RMF into two key parts, the "foundational information" and the "core and profiles". The foundational information section covers the framing of risk, the audience, AI risks and trustworthiness, and the effective use of the framework, whilst the core aims to enable dialogue, understanding, and activities to manage AI risks and to develop trustworthy and responsible AI systems. The profiles are implementations of the different AI RMF functions, as discussed below. They include temporal profiles of snapshots of current state or desired state and highlight gaps to be addressed in the AI risk management. The AI RMF profiles also include cross-sectoral profiles, for both cross-sectoral use cases and sectors.

Adapting traditional risk management and assessment definitions [25], within the foundational information NIST defines risk as a composition of an event's probability and the degree of consequences, or consequences, of it. They highlight that the consequences can be positive, negative, or both, and can give rise to opportunities or threats, and whilst risk management typically tries to address the negative consequences, in this framework NIST also try to maximize the positive consequences too.

Harms of AI systems are categorized into the harms to: i) people; ii) an organization; iii) an ecosystem.

The harms to people include harms to a person's rights, physical or psychological safety, or economic opportunity. The harms to an organization include its business operations, security, or reputation. The harms to an ecosystem include harm to interconnected and independent elements and resources, to financial or supply chain systems, or to natural resources and the environment.

NIST highlights that ill-defined or understood AI risks are difficult to quantitatively measure, however this does not mean that a given AI system does not necessarily pose a high or low risk for these. Some examples of these that NIST give are risks from third-party software, hardware, or data, emerging risks, availability of reliable metrics, and risks in real-world settings, amongst others.

Risk tolerance is the readiness of an AI actor or organisation to be a risk to achieve their objectives. The AI RMF can be used to prioritise risk but does not prescribe a risk tolerance as doing so is highly contextual, and as the tolerances are likely to change over time during the AI system lifecycle. Before risks can be prioritised and managed using the AI RMF, the risk tolerance first needs to be defined. Risk prioritisation is also contextual, with the organisation determining which risks should have the highest prioritisation for the AI system and its given use. The risk prioritisation may be different between AI systems designed or deployed to, e.g., interact with people and those which are not, or in those which are trained on large sensitive datasets or personally identifiable information, or have consequences on people. There should also be regular assessment and prioritisation of risks, and residual risks should be documented to inform users interacting with the system.

To achieve better outcomes, AI risk management should be incorporated into the broader risk management surrounding an AI system and AI risks should be considered alongside the other risks present, such as those regarding cybersecurity and privacy, rather than being considered solely on their own. As such the AI RMF can be used alongside other frameworks that aim to manage those risks, whilst it is used to manage the AI risks. Accountability mechanisms, roles and responsibilities, culture, and incentive structures are also needed by organisations for risk management to be effective, as is risk management across all stages of the AI lifecycle, with the input of a wide range of AI participants that represent diversity of experience, expertise, and backgrounds.

NIST identifies characteristics of trustworthy AI systems as: i) valid and reliable; ii) safe; iii) secure and resilient; iv) accountable and transparent; v) explainable and interpretable; vi) privacy-enhanced; vii) fair with harmful bias managed.

These AI trustworthy characteristics are classified into the classes of technical, socio-technical, and guiding principles.

Creating a trustworthy AI system requires balancing these different characteristics, with a given context of use in mind, and where enhancing some characteristics may lead to an increased compromise of others. Therefore, effective AI risk management requires balancing these trade-offs between the different trustworthiness characteristics. Additionally, through enhancing these trustworthiness characteristics, negative risks of AI can be reduced. These characteristics are socio-technical system attributes. NIST asserts that the concepts of Validity and Reliability are necessary conditions for all trustworthiness attributes. The concepts of Accountability and Transparency also relate to all the other trustworthiness attributes. As NIST points out, the assessment of these trustworthiness characteristics, the risks, consequences, costs, and benefits should inform the decision to develop or deploy a given AI system in each context.

In the context of AI systems, NIST uses the ISO definitions to define the Valid and Reliable, and Safe characteristics, with validation being the "confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled"[26], reliability being the "ability of an item to perform as required, without failure, for a given time interval, under given conditions" [27], and safety being "not under defined conditions, lead to a state in which human life, health, property, or the environment is endangered" [27]. They adapt the ISO definition of resilient [27] to define it as "can withstand unexpected adverse events or unexpected changes in their environment or use – or if they can maintain their functions and structure in the face of internal and external change and degrade safely and gracefully when necessary", and define security as "can maintain confidentiality, integrity, and availability through protection mechanisms that prevent unauthorised access and use". In the AI RMF, NIST also define transparency as "the extent to which information about an AI system and its outputs is available to individuals interacting with such a system – regardless of whether they are even aware of doing so" and state that accountability presupposes transparency and that "maintaining organization practices and governing structures for harm reduction, like risk management, can help lead to more accountable systems". NIST also define explainability as "a representation of the mechanisms underlying AI systems' operation", interpretability as "the meaning of AI systems' output in the context of their designed functional purposes", privacy as "the norms and practices that help to safeguard human autonomy, identity, and dignity", and fairness as "concerns for equality and equity by address issues such as harmful bias and discrimination". Finally, NIST identify three main categories of AI bias that should be considered and managed. These are systemic, computational and statistical, and human-cognitive bias.

The AI RMF Core is comprised of four functions (Govern, Map, Measure, Manage).

Govern

At their very highest levels, Govern is defined as "a culture of risk management is cultivated and present", Map as "context is recognized and risks related to context are identified", Measure as "identified risks are assessed, analysed, or tracked", and manage as "risks are prioritised and acted upon based on a projected consequence". Govern is cross-cutting and interwoven throughout the AI risk management process and the other functions.

In more detail, Govern makes sure that organisational policies, processes, and practices for the mapping, measuring, and managing of AI risks are present, transparent, and effective. It also makes sure that throughout the AI lifecycle accountability structures are present, that there is diversity in experience, expertise, and backgrounds in the mapping, measuring, and managing of AI risks, that AI risks are considered and communicated, that there is effective engagement with AI participants, and that policies and procedures are present for AI risks and benefits from third parties, whether that's for software, data, or other supply chain components.

Map

Map establishes the context of AI risks and provides understanding to it. It categorises the AI system and sets out to understand the capabilities, targeted usage, goals, expected benefits, and costs of it, compared to appropriate benchmarks. It also maps the risks and benefits of AI system components, and characterises the consequences to different groups, such as individuals, groups, organisations, and society. The outcomes of performing the Map function feed into the Measure and Manage functions.

Measure

The Measure function analyses and assesses AI risks and their consequences. For this qualitative, quantitative, or mixed methods can be used, and this should be performed both before the AI system deployment and throughout the AI system lifecycle. Measure includes identifying and applying appropriate methods and metrics, evaluating the AI system for trustworthy characteristics, making sure that mechanisms for tracking identified AI risks are present, and gathering and assessing feedback on the efficacy of the measurements.

Manage

The Manage function allocates resources to the mapped and measured risks and covers risk treatment plans to respond to and recover from AI risk event and communicates about them too. Manage includes prioritizing, responding to, and managing mapped and measured AI risks. It also includes the planning and implementing of strategies that maximize AI benefit and minimize the negative consequences, both from within the organization and from third parties, and also response, recovery, and communication plans.

The AI RMF gives a comprehensive non-prescriptive framework for organizations working with AI systems.

3.1.3 Comparing the AI Risk management Frameworks

Both the ENISA framework and the NIST framework use very similar definitions of an AI system, with both being based on the OECD 2019 definition, and the NIST AI RMF also being based on the ISO/IEC definition. Both consider AI systems as socio-technical systems, and both identify similar AI trustworthy characteristics and identify that there will be trade-offs when optimising one characteristic over another.

The ENISA framework is a three-layer framework which builds up on each layer, from the ICT infrastructure that an AI system is on, to the AI system itself in a sector-agnostic manner, to the AI system as used within the given sector. The NIST framework is a four-function framework that maps out, measures, and manages the risks of AI, whilst maintaining governance throughout such that there is accountability and transparency to the AI risks present in a system.

The ENISA framework categorises AI threats into eight categories based on the types of attacks, whilst the NIST framework instead categorises AI harms into three broad categories based on if the harms at the person-, organisation-, or ecosystem-level. ENISA identify eleven (11) trustworthy characteristics of an AI system whilst NIST identify seven AI trustworthy characteristics that are broader and that ENISA's eleven fit within.

[1]Both frameworks complement each other with their definitions, approaches, and differences. Both set out guidelines to work within though leave an exact and implemented methodology to the reader. As such, an implemented approach to AI risk assessment can be formed from them and be compatible with both.

Mapping between the AI Act and NIST trustworthiness characteristics

The High-Level Expert Group (HLEG) on AI[2] and the NIST Trustworthiness characteristics[3] provide complementary frameworks for ensuring the responsible development and deployment of AI systems. The AI Act draws on the ethical guidelines established by the HLEG but it remains a regulatory legal/instruments containing regulatory obligations. Ethics are broader than merely complying with the regulations. On the other hand, the NIST framework provides a technical foundation to implement trustworthiness principles in AI models. This mapping aims to align these two perspectives, demonstrating how the EU AI Act's requirements correspond to NIST's trustworthiness dimensions. The justifications below explain why each mapping is appropriate, emphasizing the synergy between regulatory oversight and technical implementation.

1. **[HLEG] Human Agency and Oversight → [NIST] Valid and Reliable**

[1] Polemi N, Praca I, Kioskli K, Becue A. Challenges and Efforts in Managing AI Trustworthiness Risks: A state of knowledge. Frontiers in Big Data. 2024, 7(1):1-14.

[2] https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf

[3] https://www.nist.gov/trustworthy-and-responsible-ai

Justification:

- The EU AI Act emphasizes the need for AI systems to support human decision-making, not replace or undermine it.
- "Valid and Reliable" in NIST ensures that AI systems consistently perform as expected across different conditions and datasets.
- This is crucial for human oversight, as AI reliability ensures that users can trust AI predictions and make informed decisions rather than being misled by inconsistent or erroneous outputs.

Note: Moreover, human agency and oversight seem to concern the interplay between human and AI which makes the human agency and oversight not to be easily mapped directly to any NIST characteristic.

2. **[HLEG] Technical Robustness and Safety → [NIST] Safe, [NIST] Valid & Reliable**

   Justification:

   - Reliability is about consistent and stable performance over time and across different conditions.
   - A technically robust AI system should behave consistently when subjected to noise, adversarial inputs, or variations in data.
   - The AI Act mandates that high-risk AI systems should be designed to prevent harm and operate securely under all foreseeable conditions.
   - "Safe" in NIST trustworthiness refers to AI systems being tested, validated, and deployed in ways that ensure safety for users, avoiding unintended harmful consequences.
   - AI failures in medical diagnostics or industrial automation can lead to physical or societal harm, making safety a key component of robustness.

3. **[HLEG] Technical Robustness and Safety → [NIST] Secure and Resilient**

   Justification:

   - The AI Act acknowledges that AI systems must withstand cyber threats, adversarial attacks, and operational failures.
   - "Secure and Resilient" in NIST emphasizes AI systems being protected from cybersecurity threats, model manipulation (adversarial attacks), and unexpected failures.

- A robust AI system is not only safe in normal operations but also capable of resisting and recovering from security breaches and environmental uncertainties.

4. **[HLEG] Privacy and Data Governance → [NIST] Privacy-enhanced**

   Justification:

   - The AI Act highlights the importance of data protection, confidentiality, and compliance with GDPR and other regulations.
   - "Privacy-enhanced" in NIST aligns with ensuring that AI systems preserve user privacy through techniques like encryption, differential privacy, and federated learning.
   - This is particularly critical in healthcare, and public sector AI applications, where personal data must be handled responsibly.

5. **[HLEG] Transparency → [NIST] Explainable & Interpretable, [NIST] Accountable & Transparent**

   Justification:

   - The AI Act mandates that AI systems should provide clear explanations for their decisions, enabling users to understand, challenge, and audit them.
   - "Explainable & Interpretable" in NIST supports this principle by ensuring that AI models are not black boxes and can provide meaningful explanations for their predictions.
   - This is essential in high-risk AI applications like healthcare diagnostics, where decision transparency is necessary for ethical and legal compliance.

6. **[HLEG] Transparency → [NIST] Accountable and Transparent**

   Justification:

   - The AI Act includes transparency as a way to hold AI systems accountable for their decisions and consequences.
   - "Accountable and Transparent" in NIST means AI models must allow for audits, tracking of decision-making processes, and clear documentation of model performance.
   - This is crucial for regulators, companies, and end-users who need to ensure AI is functioning as intended and does not produce unfair or harmful outcomes.

7. **[HLEG] Diversity, Non-Discrimination, and Fairness → [NIST] Fair-with harmful bias managed**

Justification:

- Regulations at EU and national level, such as the AI act or anti-discrimination laws explicitly requires AI systems to be free from unjust bias, ensuring they do not disadvantage specific demographic groups.
- "Fair-with harmful bias managed" in NIST refers to AI models actively identifying, mitigating, and reducing harmful biases that could consequence fairness.
- AI-driven law enforcement applications must be fair and not reinforce societal inequalities.

8. **[HLEG] Societal and Environmental Well-Being → [NIST] Safe**

   Justification:

   - The AI Act extends AI trustworthiness to include its consequence on society and the environment, ensuring AI systems contribute positively and do not pose risks.
   - "Safe" in NIST aligns with this by ensuring AI technologies are developed responsibly, reducing risks to both individuals and the broader society.
   - Examples include AI systems in industrial automation, healthcare AI-based systems or other examples of critical domains, where poorly designed AI could lead to negative societal and environmental consequences.

   Note: Societal and environmental well-being in AI Act seems to address something far broader than the NIST safe thus societal and environmental well-being seems more accentuated by HLEG than by NIST.

9. **[HLEG] Accountability → [NIST] Accountable and Transparent**

   Justification:

   - The AI Act enforces accountability by ensuring that high-risk AI systems have clear responsibility structures, logs, and mechanisms to trace errors back to developers or operators.
   - "Accountable and Transparent" in NIST supports this by requiring that AI decision-making processes are well-documented, explainable, and auditable.
   - For example, this is critical in healthcare diagnostics, where accountability ensures AI decisions are fair, lawful, and contestable.

3.2 Landscape of standards

The landscape of standards surrounding AI risks is vast, encompassing contributions from ISO/IEC, ETSI, and IEEE (IEEE P2976, 2021; IEEE P3119, 2021). This section highlights key standards that focus on AI risks and trust management. Starting with the foundational standards for risk management, such as the ISO27000x series and ISO 31000:2018, we then move to dedicated AI risk management standards like ISO/IEC 24028, which addresses AI security threats. ISO/IEC 42001—Artificial Intelligence Management System, published in December 2023, is designed to manage risks and opportunities associated with AI, addressing ethics, transparency, reliability, and continuous learning. ISO/IEC 23894 works in conjunction with ISO 31000:2018, focusing specifically on AI risk management. ISO/IEC has also published TR standards, including those that concentrate on AI ethical and societal concerns.

The robustness of neural networks is tackled by ISO/IEC 24029-2:2023, which offers a methodology for using formal methods to assess neural network robustness. The development of ISO/IEC 24029- 3 aims to focus on statistical methods for this purpose. Technical Report TR 24028 analyses and surveys approaches to enhance trustworthiness in AI systems and mitigate vulnerabilities related to trustworthiness. Other relevant ISO standards include:

• ISO/IEC WD 27090—Cybersecurity—Artificial Intelligence: Guidance for addressing security threats to AI systems.

• ISO/IEC WD 27091—Cybersecurity and Privacy: Artificial

Intelligence—Privacy protection.

• ISO/IEC 27115—Cybersecurity evaluation of complex systems: Introduction and framework overview.

• ISO/IEC CD TR 27563: Consequence of security and privacy in AI use cases.

• ISO/IEC 5338 (also covering the AI risk management process and summarizing 23894).

• ISO/IEC AWI 42105 (under development) on guidance for human oversight of AI systems.

• ISO/IEC 5259 series (Data quality).

• ISO/IEC 24029 series (Robustness).

• ISO/IEC 22989 (AI concept and terminology standard).

• ISO/IEC FDIS 5338: AI system lifecycle processes.

From ETSI, the Securing Artificial Intelligence (SAI) group is making strides in this area. It published the AI Threat Ontology [ETSI GR SAI 001 V1.1.1 (2022-01)] as one of its initial reports. In 2023, ETSI introduced the Artificial Intelligence Computing Platform Security Framework [ETSI GR SAI 009 V1.1.1 (2023- 02)], detailing a security framework for AI computing platforms to protect valuable assets like models and data. Additionally, ETSI GR SAI 007 V1.1.1 (2023-03) discusses steps for AI platform designers and implementers to ensure explicability and transparency in AI processing.

IEEE has introduced P3119, a standard for the Procurement of Artificial Intelligence and Automated Decision Systems, establishing definitions and a process model for AI procurement and how government entities can address socio-technical and innovation considerations responsibly. The IEEE P2976—Standard for XAI (eXplainable Artificial Intelligence)—aims to define the requirements for an AI method, algorithm, application, or system to be considered explainable, ensuring clarity and interoperability in AI system design.

In March 2023, the European Commission (EC) requested CEN and CENELEC to work with international and national stakeholders, including SMEs, to develop a European standards program for AI (CEN/CENELEC Standards, 2023). These standards will aim to ensure safety, transparency, user understanding, oversight, accuracy, robustness, cybersecurity, and quality management throughout the AI systems' lifecycle, catering to various stakeholders' needs and ensuring regulatory compliance. This request by the EC was accompanied by a set of requirements in the following areas for the new EU standards:

Risk management system for AI systems: Specifies a continuous iterative process for risk management throughout the AI system's lifecycle, aimed at preventing or minimizing risks to health, safety, or fundamental rights. Ensures integration of risk management systems with existing Union Harmonization legislation where applicable. Drafted for usability by relevant operators and market surveillance authorities.

Data and data governance: Includes specifications for data governance procedures, focusing on data generation, biases, and dataset quality for training AI systems.

Record keeping through logging capabilities: Specifies automatic logging of events for traceability and post-market monitoring of AI systems by providers.

Transparency and information to users: Provides design and development solutions for transparent AI system operations and instructions for users about system capabilities and limitations.

Human oversight: Specifies measures and procedures for human oversight built into AI systems and implemented by users, including those specific to certain AI systems' intended purposes.

Accuracy specifications for AI systems: Lays down specifications for ensuring appropriate accuracy levels, declaring relevant accuracy metrics and tools for measurement.

Robustness specifications for AI systems: Specifies robustness considering potential sources of errors, faults, and interactions with the environment.

Cybersecurity specifications for AI systems: Provides organizational and technical solutions to safeguard AI systems against cyber threats and vulnerabilities.

Quality management system for providers of AI systems: Specifies a quality management system ensuring continuous compliance with various AI system aspects.

Conformity assessment for AI systems: Provides procedures for conformity assessment activities related to AI systems and quality management systems of AI providers. Another area of development for standards and methodologies is that of General Purpose AI (GPAI).

3.3 Legislation and Policies

This section aims to elucidate several EU legislative instruments and policies concerning Trustworthy AI. First, it highlights the well-recognised ethical principles of AI trustworthiness established by the EU High-Level Expert Group (HLEG) at European level. Second, it examines relevant binding legal instruments concerning the regulation of AI systems with a particular focus on fundamental rights, data protection rules and the AI Act. Those legal instruments adopt a risk-based approach and mandate risk assessments under specific conditions to ensure the responsible use of AI systems complementing trustworthy AI principles, and should be taken into account for the development and use of AI systems during FAITH project and afterwards.

The presence of legal, ethical and societal risks associated with AI systems necessitates the evaluation of the legal and ethical adherence of Trustworthy AI (EC HLEG, 2019; EC, 2023). From a legal perspective, the role of risk in legislation is twofold. On the first side, the law itself targets risks arising from the specific process or technology, including AI systems, in an identified domain or context, and depending on the risk level, the legislations put forward strict requirements to be complied with by stakeholders to avoid those risks from being realised. On the other side, non-compliance by rules enshrined in legislation itself is a risk factor for stakeholders that should be minimised to avoid both risks targeted in law to be realised and to avoid fines and punishments. With that said two important EU regulations concerning AI technologies which rely on risk-based regulatory approaches are the GDPR and the AI Act.

The focus of section concerns:

1) FAITH Trustworthiness Evaluation Technical Infrastructure – FAITH STM (‘FAITH tool’) *per se* needs to abide by the legal and ethical requirements for Trustworthy AI, given that it is an AI-powered technology. The legal and ethical consequence assessment of the FAITH tool will be carried out in WP 1 T1.4 and will be analysed in deliverables D1.3 and D1.4.

2) The FAITH tool will analyse the risks of third-party AI systems and recommend points of action for their better comply with trustworthiness requirements, tailoring them to their domain understanding and implementation while ensuring legal and ethical compliance.

3.3.1 AI HLEG principles

The High-Level Expert Group on Artificial Intelligence (HLEG) was established by the EC with the aim of providing specific guidance on AI strategies. The HLEG has produced a landmark document to promote the ethical use of AI: the "Ethics Guidelines for Trustworthy AI". The HLEG's Ethics Guidelines for Trustworthy AI is examined in this section, as it is very relevant to the technologies developed within the FAITH Project.

Trustworthy AI encompasses three criteria that should be met throughout the system's entire life cycle

- Being **lawful, complying with all applicable laws and regulations**;
- Being **ethical, ensuring adherence to ethical principles and values**; and
- Being **robust, both from a technical and societal perspective, since, even with good intentions, AI systems can cause unintentional harm**.

The Guidelines outline seven fundamental principles for the development of AI systems: **1) human agency and oversight; 2) technical robustness and safety; 3) privacy and data governance; 4) transparency; 5) diversity, non-discrimination and fairness; 6) societal and environmental well-being; and 7) accountability**.

To assist technology developers and businesses in the implementation of these principles, the HLEG has developed the "Assessment List for Trustworthy AI" (ALTAI) [28], available both as a document and as a prototype of a web-based tool.

3.3.2 Fundamental Rights

In the EU, fundamental rights and freedoms are recognised and protected by several legal documents on international, supranational (EU) and national level. Among others, one of the key documents is the EU Charter on Fundamental rights (CFREU, the EU Charter), which establishes and guarantees fundamental rights in the EU and across EU member states. Similar fundamental rights are also guaranteed under national constitutions. Regarding the digital environment, the EU issued in 2023 the European Declaration on Digital Rights, a non-binding document.

Although fundamental rights are not absolute and can be limited, those limitations must be legal and must respect the essence of rights and freedoms. Actions and technologies affecting fundamental rights must be necessary in order to attain a legitimate aim, and the chosen practice must be the least intrusive method to achieve such aim. These aspects must be considered when evaluating the legality of a certain process or technology, including FAITH project activities, following what is known as the necessity and proportionality assessment. To name a few, regarding the use of AI systems, the right to privacy, right to data protection, right to non-discrimination and equality is important when designing, developing and using AI- powered technologies under the FAITH Project. Hence, it is crucial that all activities, including those carried out throughout the FAITH research and deployment phases, strictly adhere to these fundamental rights. In this regard noncompliance with those rights can be considered as an important risk factor, and this should be analysed both for the AI- powered FAITH tools and AI systems used in each FAITH pilot.

Hence, it is important to evaluate fundamental rights risks that may result from the AI-powered FAITH tool and the AI systems used in the LSPs, taking into account the specific applications, determining adequate mitigation measures and state-of-the-art best practices to minimise these risks (including meaningful human intervention), and meticulously implement them during the research and operational stage to avoid any harm.

3.3.3 GDPR

EU citizens are granted the rights to privacy and data protection by the Charter of Fundamental Rights of the EU. Article 7 states that “everyone has the right respect for private and family life, home and communications”, whereas Article 8 regulates that “everyone has the right to the protection of personal data concerning him or her” and that processing of such data must be “on the basis of the consent of the person concerned or some other legitimate basis laid down by law.” These rights are developed in detail by the General Data Protection Regulation (GDPR), Regulation 2016/679/EC, which is in force in every Member State since the 25th of May 2018.

The European Union's General Data Protection Regulation (GDPR) stands as a cornerstone of data privacy in the EU. Its reach extends to the realm of scientific research as well, influencing programs like Horizon Europe, the EU's flagship initiative for funding research and innovation. This section delves into the analytical exploration of how GDPR regulations are applied to the FAITH project.

The GDPR aims to further protect the personal data of individuals and their free movement within the EU. The GDPR applies to all entities that are either fully established in the EU or have branches established in the EU that process personal data as part of their activities, regardless of where the data is processed. It also applies to entities established outside of the EU, which offer goods/services to individuals in the EU or monitor the behaviour of such individuals within the EU.

Therefore, since the 25th of May 2018, not only applicants, beneficiaries, contractors or subcontractors receiving funding from EU programmes such as H2020, but also trainers and experts, must comply with the GDPR. Any natural or legal person who collects or in any way uses for professional purposes the personal data of individuals must comply with the rules, as is the case with any other EU or national legislation they are subject to.

The GDPR applies only to the processing of personal data. Since the EU data protection legislation only deals with the processing of personal data, the distinction of personal and non-personal data (which includes anonymous data) is crucial for all activities of the project. Article 4(1) GDPR defines personal data as:

> *"any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person."*

Personal data, as defined by the GDPR, is any information related to an identified or identifiable natural person, i.e., names, identification numbers, emails, postal addresses, phone, location data, pictures, signatures, etc. This excludes information about companies, anonymised or statistical data, which is not personal data. Processing means any operation performed on the personal data, such as collecting, recording, storing, organising, filing, using, combining, disclosing, transferring, or erasing manually or automatically, i.e., collecting contact details of participants to an event, sending newsletters by email, publication of participants lists or pictures with persons related to an event, subscription to e-services etc. The GDPR establishes a set of principles and requirements that the FAITH consortium shall comply with as also promised in the FAITH consortium Agreement (Section 4.5) and FAITH Grant Agreement. FAITH Partners will comply with all the requirements specified by the GDPR to ensure full respect for the principles relating to the processing of personal data. In particular, the project shall adhere to the data protection principles of:

- **Lawfulness, fairness and transparency:**
  - In accordance with these principles, data must be processed with respect to the law, proportionally to the aim foreseen and transparently towards the data subjects concerned. FAITH will process personal data in compliance with the GDPR and other national or European applicable legislation that applies in the context of the project. FAITH will also process data fairly by balancing the data processing needs of the consortium and the rights and interests of the data subjects. FAITH shall also process data in a transparent manner, by providing information to the natural persons concerned about the collection, use and storage of their data as well as the extent of these operations, following the informed consent procedures decided by FAITH Partners. All research processes and procedures will be transparent to all stakeholders.
- **Purpose limitation:**
  - The collection and processing of personal data should be limited to specified, explicit and legitimate purposes. Following this principle, FAITH will take appropriate technical and organisational measures to ensure that, by default, only personal data which are relevant to the envisaged research are collected and processed. Personal data will be used in FAITH to pursue research objectives and tasks indicated in the FAITH Consortium and Grant Agreements. Among others, this also includes:

    - To disseminate FAITH results to stakeholders
    - To communicate news and information about the project to the public, the media and civil society organisations;
    - To support stakeholders in the exploitation of FAITH results.

- **Data minimisation:**
    - This principle entails the need for FAITH partners to ensure that personal data being processed is adequate (i.e., sufficient to properly fulfil the stated purposes of the project), relevant (i.e. the personal data has a rational link to FAITH research purposes) and limited to what is necessary (i.e. FAITH partners shall not hold more than what is necessary for the purposes of the research). FAITH shall take appropriate technical and organisational measures to ensure that, by default, only personal data which are relevant to the envisaged research are processed. In compliance with the data minimisation principle, FAITH will assess whether the same purposes can be achieved by collecting less data than initially intended and, where that is the case, apply the narrower collection option available.
- **Accuracy:**
    - Personal data shall be accurate and, where necessary, kept up to date. In accordance with this principle, FAITH will take every reasonable step to ensure that the data being processed is accurate and kept up to date. Accordingly, and having regard to the purposes for which the data are processed, FAITH partners shall erase or rectify data without delay.
- **Storage limitation:**
    - In line with this principle, FAITH shall not keep personal data for longer than is necessary for the purposes of the project. To address this, FAITH will ensure that personal data from volunteers is kept for as long as necessary and, where appropriate, in an anonymised or pseudonymised manner. Personal data collected during the FAITH research will be deleted incline with the agreed storage duration.
- **Integrity and confidentiality:**
    - According to this principle, FAITH partners (as data controllers of the processing) must have appropriate security measures in place to protect the personal data held. Once the personal data has been used for its intended purposes within the project, it will be deleted to avoid accidental risk of future disclosure (unless required to be kept for legal or contractual purposes). Such measures and procedures are intended to safeguard ethical values, such as protecting the rights and autonomy of individuals to the fullest.
- **Accountability:**
    - Accountability is the grounding principle of the GDPR. It requires the controller to adhere to and demonstrate compliance with the Regulation and the above principles. Accountability in the context of FAITH may translate, for example, into carrying out the necessary evaluations about legal and ethical procedures for FAITH research activities entailing the use

of personal data. Accordingly, FAITH partners will ensure the accountability and responsible handling of the personal data processed within the FAITH Project.

3.3.4 AI Act

The Artificial Intelligence Act (AIA) is a European Union regulation on artificial intelligence (AI) which was proposed by the European Commission on 21 April 2021 [29] and approved by the European Union's parliament and the EU Council. and will be published in the Official Journal in July 2024.

The AIA came into force on the twentieth day after its publication in the Official Journal and shall become applicable 24 months from the date of entry into force (Article 113 AI Act), except for some specific provisions. In other words, some requirements foreseen by AIA must be applied sooner or later than 24 months. For example, Chapter II on Prohibited AI Practices and provisions on AI literacy apply since 2 February 2025, whereas obligations for some high-risk AI systems that are not prescribed in Annex III but are intended to be used as a safety component of a product shall apply from 2 August 2027 (Article 113 AI Act).

The AIA sets several crucial objectives: (1) to enhance and foster the proper functioning of the single market for AI systems by setting harmonised rules in the EU, (2) to ensure a high level of **protection of health, safety, and fundamental rights** enforcing democratic values and environmental protection, and (3) to support innovation by foreseeing measures with a particular focus on SMEs.

The AI Act is a recent development. On 4 and 6 February 2025, the European Commission released two series of guidelines relevant to the AI Act: the Guidelines on prohibited AI practices and the Guidelines on the definition of an AI system.

3.3.4.1 The risk-based approach

The AI Act introduces a risk-based approach to the regulation of AI. To that end, the AI Act distinguishes between AI systems posing (1) prohibited practices, (2) high risk, (3) limited risk, and (4) low or minimal risk (see above for a visualisation). The aim is to adapt the level of obligations to the risk level of the AI system to ensure adequate protection in the EU while not deterring innovation.

3.3.4.2 Definition of AI systems

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

The EC Guidelines on the definition of an AI system highlights that the definition comprises seven elements (table 8):

Table 8: AI System elements

Machine-based System	· AI systems are developed and run on machines, including both the hardware and software components enabling the AI system to function.
Autonomy	· AI systems are designed to operate with 'some degrees of independence of actions from human involvement and of capabilities to operate without human intervention. · Exclusion of systems that are designed to operate solely with full manual human involvement and interaction. · Necessary condition to determine whether a system qualifies as an AI system.
Adaptiveness	· Self-learning capabilities allowing the behaviour of the system to change while in use (Recital 12 AI Act). · Facultative, not decisive condition for determining whether the system qualifies as AI.
AI system objectives	· Might be explicit or implicit. · They might be different from the intended purpose of the AI system in a specific context (Recital 12 AI Act), e.g. 'the use for which an AI system is intended by the provider (Article 3 (12) AI Act). · Objectives are internal to the system, while the intended purpose is externally oriented and includes the context in which the system is designed to be deployed and how it must be operated.

Inferencing on how to generate outputs using AI techniques	· Excludes systems that are based on the rules defined solely by natural persons to automatically execute operations (Recital 12 Act). · Refers to the process of obtaining the outputs, (…), which can influence physical and virtual environments, and to a capability of AI systems to derive models or algorithms, or both, from inputs or data (Recital 12 AI Act). · Technique-enabling inferences include machine learning approaches (such as supervisor learning, unsupervised learning, self-supervised learning and reinforcement learning) and logic- and knowledge-based approaches (Recital 12 AI Act).
Outputs than can influence physical or virtual environments	· Outputs of AI systems belong to four categories, all differing in the level of human involvement: predictions, content, recommendations or decisions.
Interaction with the environment	· AI systems are not passive; they actively impact the environments in which they are deployed. · Refers both to physical environments (tangible, physical objects) and to virtual environments (digital spaces, data flows and software ecosystems).

The EC Guidelines further specifies the text of Recital 12 that excludes from the definition of AI systems ‘simpler traditional software systems or programming approaches (…) that are based on the rules defined solely by natural persons to automatically execute operations’. This includes systems for improving mathematical optimization, basic data processing, systems based on classical heuristics and simple prediction systems.

3.3.4.3 Scope of application

The AIA will apply to the providers of AI systems placing or putting into service AI systems on the EU market, irrespective of their place of establishment (Article 2(1)(a) AI Act). According to the Article 3(3) AIA, a provider means ‘a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge’. The AI Act will also apply to deployers of AI systems; in other words to the ‘users of an AI system' (Article 3(4) AI Act). Under the scope of the AIA, developers and deployers hold different obligations to fulfil. The text extends its scope to apply to importers and distributors of AI systems within the EU and to product manufacturers who place an AI system on the market or put it into service with their product and under their trademark. AIA also identified other participants; **authorised representatives of providers**, which are not established in the Union; **affected persons** that are in the Union (Article 2(1)(d) and 2(1)(e) AI Act).

The AIA has put forward crucial exceptions to the scope of the AIA for research and development into AI systems. The final consolidated text provides a general exception for all the AI systems and models, including their output, specifically developed and put into service **for the sole purpose of research and development** (Article 2(6) AI Act). This exception aims to support innovation and respect freedom of science. It ensures that the AIA does not hinder scientific research and development activity on AI systems or models **prior to being placed on the market or put into service (recital 25 AI Act)**. However, **once those systems and models are put into service or placed on the market**, they must be compliant with the AIA.

Since it is difficult to ensure compliance retroactively, it is crucial to consider the AIA's requirements while conducting research on products that may be placed on the market or put into service.

If AI systems and models are not placed on the market or put into service, their testing and development activities are also exempted from the AIA. Another important point is if **the testing happens in real-world conditions**, (Article 2(8) AI Act) in such a case, the specific conditions under the AIA become applicable.

Therefore, **even though the AIA may not be applicable to the research and development phase; particular attention should be given if the testing of research is performed in real-world conditions. Furthermore, it is relevant to pay close attention to respecting the AI Act in order to create a future proof technology, especially if entry to the market is considered at later stages.**

In any event, any research and development activity should be conducted in accordance with applicable Union law and the recognised ethical and professional standards for scientific research.

3.3.4.4 Prohibited AI systems

Unacceptable risks relate to the use of AI systems considered harmful to people's safety, health and fundamental rights. As the establishment of mitigating measures would not be sufficient to attain an acceptable level of risk, those systems are prohibited in the EU. Article 5 of the AIA identifies 8 prohibited practices [30] and the Guidelines on prohibited practices provides practical guidance on the identification of these practices. These AI practices are as follow:

Table 9: AI practices

Harmful manipulation, and deception	AI systems that deploy 'subliminal techniques' to harmfully manipulate a person's behaviour. These are AI systems that use subliminal techniques that are beyond a person's consciousness with the purpose of distorting human behaviour in a way likely to cause that person or another person harm, whether physical or psychological (Article 5(1)(a) AI Act).
Harmful exploitation of vulnerabilities	AI systems that harmfully exploit the vulnerabilities of a specific given group of people (due to their age, physical or mental disability, or a specific social or economic situation) (Article 5(1)(b) AI Act).
Social scoring	AI systems that harmfully exploit the vulnerabilities of a specific given group of people (due to their age, physical or mental disability, or a specific social or economic situation) (Article 5(1)(b) AI Act).
Individual criminal offence risk assessment and prediction	AI systems for making risk assessments of natural persons in order to predict the risk of a person committing a criminal offence, based solely on the profiling of a person or on assessing their personality traits. On the other hand, this prohibition does not extend to AI systems used to support the human assessment of the involvement of a person in a criminal activity, which is already based on objective and verifiable facts directly linked to a criminal activity (Article 5(1)(d) AI Act).
Untargeted scraping to develop facial recognition databases	AI systems creating or expanding a facial recognition database through the untargeted scraping of facial images from the internet or CCTV footages (Article 5(1)(e) AI Act).
Emotion recognition	AI systems inferring emotions to be used in workplaces and education institutions (Article 5(1)(f) AI Act).
Biometric categorisation	Biometric categorisation systems used to infer a person's sensitive personal data (i.e., race, political opinion, trade union membership, religious or philosophical belief, sex life or sexual orientation). However, this prohibition does not include any labelling or filtering of lawfully acquired biometric datasets, or

	categorizing of biometric data in the area of law enforcement (Article 5(1)(g) AI Act).
Real-time remote biometric identification (‘RBI’)	‘Real-time’ remote biometric identification systems (e.g. facial recognition) in publicly accessible spaces for law enforcement purposes, subject to specific exemptions (Article 5(2)(3) and (4) AI Act).

3.3.4.6 High-risk AI systems

High-risk AI systems are those that could negatively consequence individuals’ health, safety and fundamental rights. However, their use is permitted if they comply with the obligations set forth in the AI Act and have minimized associated risks. The original AI Act proposal foresaw two categories of such systems, based on their purpose and use:

1. AI systems used as a safety component of a product or falling under EU health and safety harmonisation legislation (such as toys, cars, aviation, medical devices, lifts, as stipulated under the Annex I) (Article 6(1)(a) AI Act).
2. AI systems deployed in any of the eight specific areas identified in Annex III (1. biometric identification and categorisation of natural persons; 2. management and operation of critical infrastructure; 3. education and vocational training; 4. Employment, workers management and access to self-employment; 5. access to and enjoyment of essential private services and public services and benefits; 6. law enforcement; 7. migration, asylum and border management; 8. and administration of justice and democratic processes).

The EC can adopt delegated acts to amend Annex III (Article 7(1) AI Act). In conducting this evaluation, the EC needs to consider a wide range of factors, including the AI system’s intended purposes and extent of use, the nature and amount of personal data processed, the possible human oversight, imbalances of power, the likelihood of benefits for individuals (Article 7(2) AI Act).

3.3.4.7 Requirements for High-Risk AI systems

All providers of high-risk AI systems will be subjected to a set of legal requirements listed in Articles 8 and 9 of the AI Act.

One of the core requirements for the high-risk AI systems is the establishment, implementation, documentation and maintenance of a risk management system. This risk-management systems should consist of four steps:

(1) identify and analyse the known and foreseeable risks; (2) estimate and evaluate the risks that may emerge from normal use and foreseeable misuse; (3) evaluate other possible risks,

based on information collected from the post-market monitoring system; and (4) adopt suitable risk management measures, taking into account the state of the art (Article 9(4) AI Act). Any residual risks after application of the risk management measures must be communicated to the user (Article 9(5) AI Act). Thus, a risk management system shall consist of 'a continuous iterative process throughout the AI system lifecycle' and it needs to be updated systematically.

High-risk AI systems must fulfil high data quality and data governance standards (Article 10 AI Act)**.** Data governance standards include relevant design choices', 'data preparation', and prior assessment of the datasets and examination in view of possible biases likely to affect the fundamental rights, health and safety of the subjected persons or to lead to prohibited discrimination (Article 10(2) AI Act). Hence, identification of potential data gaps or shortcomings with plans to address is a crucial requirement. Training, validation and testing datasets must be 'relevant, sufficiently representative and, to the best extent possible, free of errors and complete' (Article 10(3) AI Act).

Providers can **exceptionally process special categories of personal data**, as defined under article 9 of the GDPR, in so far as this is strictly necessary to effectively detect and correct biases (Article 10(5) AI Act). The special categories of personal data processed in this context cannot be re-used or transmitted to other parties. Such data need to be deleted once the bias has been corrected (Article 10(5)(e) AI Act). Furthermore, if special categories of data are processed, safeguards must be implemented to protect the rights to privacy, data protection and other fundamental rights. To that end, pseudonymization, encryption of data, or other technical limitations on re-use and security measures shall be enforced. Strict controls and documentation of the access are necessary to avoid misuse and ensure that only authorised persons access such data (Article 10(2) (b) and (c) AI Act).

Another crucial requirement for high-risk AI systems is the obligation to ensure a sufficient degree of transparency to enable users to interpret their outputs and use it appropriately (Article 13 AI Act)**.** In that regard, the contact details of the provider, the characteristics, capabilities and limitations of performance of the high-risk AI system, the changes to the high-risk AI system and its performance and the human oversight measures referred to in Article 14 must be shared. The transparency extends to exposing the computational and hardware resources, as any necessary maintenance and care measures, including their frequency, to ensure the proper functioning of that AI system. Hence, it would be possible to detect the expected lifetime of the high-risk AI system and to ensure its proper functioning (Article 13(3) AI Act).

The technical documentation is also essential, and it must be drawn up before high-risk AI systems are placed on the market or put into service. The documentation must be constantly updated. The technical documentation shall demonstrate the compliance with the high-risk AI system with the necessary, clear and comprehensive information to enable the competent authorities to assess the compliance (Article 11(1) AI Act).

The high-risk AI system must include automatic logging functions that allows for traceability of the system's functioning over its lifecycle. The system must also be designed in a way that ensures that operation is sufficiently transparent to allow users to interpret its output and use the system appropriately (Article 12 AI Act).

Furthermore, **high-risk AI systems need to guarantee effective human oversight** (Article 14 AI Act)**.** Human-machine interface shall allow individuals to understand and oversee the high-risk AI system. Human oversight is to prevent or minimize risks to health, safety, or fundamental rights. It provides a view of the capabilities and limitations of the system and allows for intervention using a "stop" button or other similar procedures (Article 14(4) AI Act). Particularly for biometric identification AI systems described in Annex III, point 1(a) of the AI Act, human oversight measures should ensure that no action or decision should be taken on the identification, unless at least two people with the necessary competence, training and authority verify and confirm such decision (Article 14(5) AI Act).

As part of **the requirements for accuracy, robustness and cybersecurity**, the high-risk AI systems shall be developed in a way to ensure that feedback loops are adequately addressed with mitigation measures. Accuracy, robustness and cybersecurity should then be maintained throughout the life cycle of the AI systems (Article 15 AI Act).

Finally, providers of high-risk AI systems must ensure that a conformity assessment is carried out prior to placing the system on the market or putting it into service. High-risk AI systems will be subject to a conformity assessment through already existing conformity frameworks and harmonised standards (e.g. for medical devices) or through the conformity assessment procedure based on internal control (in line with the Annex VI), or through the conformity assessment procedure based on assessment of the quality management system and technical documentation, with the involvement of a notified body (in line with the Annex VII).

In case the assessment shows that the requirements foreseen in the AI Act have been met, as stipulated under Article 16 AI Act, the providers shall draw up an EU declaration of conformity and attach the CE marking of conformity. According to the Article 23 AI Act, the importers who place a high-risk AI system on the market must also ensure that the provider has done the conformity assessment, affixed the conformity marking, and included the required documentation and instructions for use [31].

Finally, according to the Article 27 AI Act, public entities (e.g., municipalities, public administrative bodies) **shall conduct a fundamental rights impact assessment of their high-risk AI systems**, prior to their deployment. The AI Office will develop a template and an automated tool to facilitate compliance with the obligation as stated under Article 27(5) AI Act. A detailed analysis of impact assessments including Fundamental Rights impact Assessment will be provided in Deliverable 1.3.

3.3.4.8 Requirements for Limited, low or minimal risk AI Systems

Compared to high-risk AI, a smaller section of the AI Act handles limited-risk AI systems. **Such systems are subject to more limited transparency obligations** as foreseen under Article 50 'Transparency obligations for providers and deployers of certain AI systems.' For example, **AI systems communicating with humans as chatbots** or AI systems that generate deepfakes can be considered limited-risk AI systems. In the case of using such AI systems, **in line with the transparency requirements stated under the** Article 50(4) AI Act**, individuals must be warned that they are interacting with an AI system**.

Deployers who use an AI system to generate deep fakes should also clearly and distinguishably disclose that **the content has been artificially created** or **manipulated by labelling the AI output accordingly and disclosing its artificial origin**. On the other hand, this transparency obligation shall not hamper the right to freedom of expression and the right to freedom of the arts and sciences, in particular where the content is part of an evidently creative, satirical, artistic, fictional, or analogous work or programme. In addition, it is also appropriate to envisage **a similar disclosure obligation about AI-generated or manipulated text** published to inform the public on matters of public interest unless the AI-generated content was reviewed by a human or went through an editorial control and a natural or legal person holds editorial responsibility for the publication of the content as stipulated under the recital 134 AI Act.

3.3.5 Non-discrimination Laws

In the European Union, there is not one comprehensive piece of legislation concerning protection against discrimination. While companies can consult the General Data Protection Regulation (GDPR) for most answers related to the lawful processing of personal data, they face a more fragmented regulatory framework when dealing with non-discrimination rights. Two key considerations should be considered by FAITH when acting to comply with non-discrimination laws in the European Union.

Firstly, the level of protection against discrimination is partially harmonized by European Union law across all 27 EU Member States. This means that when FAITH Partners examine the EU Equality Legal Framework, they can determine the minimum level of protection against discrimination they must comply with, whether operating in Spain or in Greece, for instance.

Secondly, while each of the 27 Member States can enhance the level of protection in their national laws, they cannot reduce it below EU standards. In practical terms, when FAITH Partners review individual EU Member States' national legislation, they will discover additional rules they must comply with beyond the EU Equality Legal Framework.

We can illustrate this situation with an example. The EU Legality Framework provides that it is illegal discrimination the differentiation based on sex and race in the access to and supply of goods and services available to the public. In the context of one of FAITH LSPs, being able to access public transportation refers to access to services. According to the EU Equality Legal Framework, any AI tool governing this access cannot base its decision on aspects protected against discrimination. Furthermore, national anti-discrimination law might extend this protection in relevant EU countries.

At this stage of the FAITH research project, a sensible starting point for developing an anti-discrimination strategy would be to consider the EU Equality Framework as a baseline. This framework is consistently reflected across all 27 EU Member States' laws. However, for FAITH to develop tools that complies with all European Union countries' non-discrimination laws, it must account for the protected characteristics specified by each of the 27 Member States regarding equal treatment in goods and services access.

It is important to note that providers and deployers of AI tools should not only address direct discrimination but also indirect discrimination, intersectional discrimination, discrimination by association and discrimination by perception.

3.3.6 Cybersecurity Regulation

Cybersecurity obligations in the EU apply to AI systems as they are cyber assets within an ICT infrastructure, composed of different AI assets such as data, models, processes and tools. This involves cybersecurity consequence assessments. FAITH tools as well as LSPs should respect obligations within this framework, notably to ensure AI robustness.

3.3.6.1 NIS Directives and NIS II

The NIS 2 Directive [32], set to replace NIS 1 by October 18, 2024, strengthens existing cybersecurity rules for critical sectors across the EU. While member states had until 17th October 2024 to implement the directive, currently only 5 member states transposed the directive [33].

Article 2 and 3 of the Directive and its Annexes designates critical sectors such **as waste water, transport and healthcare** for which member states must identify entities. The directive further categorizes entities as either "essential" or "important" based primarily on size criteria, with exceptions for highly critical operators.

These entities must implement appropriate technical, operational, and organizational measures to manage risks to their network and information systems, including AI systems. The measures must consider state-of-the-art technology, relevant standards, implementation costs, and the entity's risk exposure, while taking an all-hazards approach that addresses **physical, environmental, human, and interference risks**[4].

[4] See https://eur-lex.europa.eu/legal-content/FR/NIM/?uri=CELEX:32022L2555, consulted on 23 October 2024

3.3.7 Cybersecurity Act

The Cybersecurity Act (CSA)[5] introduced a comprehensive certification framework for ICT products, services, and processes, including AI technologies. While certification remains voluntary, it provides a presumption of compliance with cybersecurity requirements.

A **cybersecurity certification scheme** is **"a comprehensive set of rules, technical requirements, standards and procedures that are established at Union level"** and which serves to assess the cybersecurity of specific ICT products, ICT services or ICT processes.

Current security assessments for AI systems primarily rely on existing standards and methodologies not specifically designed for AI technologies [33]. According to the Article 42(2) AI Act, high-risk AI systems falling under a cybersecurity scheme under the CSA are presumed to comply with the cybersecurity requirements posed by article 15 of the AI Act.

3.3.8 Cyber Resilience Act

The upcoming Cyber Resilience Act (CRA) [34] introduces common cybersecurity requirements for **products with digital elements**, encompassing both software and hardware. AI systems are encompassed in the definition of products with digital elements. Article 10(1) and (2) CRA states that the regulation ensures that product with digital elements whose intended or reasonably foreseeable use includes a connection with a device or a network must undergo a **cybersecurity risk assessment** before being put into the market, with a view to minimising cybersecurity risks, preventing security incidents and minimising the consequences of such incidents, including in relation to the health and safety of users. Moreover, the Article 13(3) CRA states that the assessment consists of an analysis of cybersecurity risks and an indication of the way the cybersecurity requirements are implemented in practice. This assessment starts at the design phase and continues throughout development and production as stipulated under the Article 13(3)(8) CRA. According to the Article 12 CRA, high-risk AI systems that meets the requirements of the CRA are automatically compliant with the requirements posed by article 15 of the AI Act.

3.4 AI Technologies

According to the OECD [35] definition, an AI system operates as a machine-based entity capable of influencing its surroundings by generating outputs—such as predictions, recommendations, or decisions—that align with specific objectives. It leverages machine-generated and human-provided data and inputs to: (i) perceive and interpret real or virtual environments; (ii) abstract these interpretations into models, using automated analysis (e.g., machine learning) or manual methods; and (iii) employ model inference to generate potential outcomes. AI systems are designed to function with varying levels of autonomy.

AI encompasses a broad spectrum of disciplines, each of which can be further subdivided into various specialized fields that are sometimes used interchangeably. Here are some examples:

[5] NIS 2, Art. 21

Computer vision: This field involves the automatic processing of visually rich data such as images and videos. Key tasks within computer vision include object detection, facial recognition, action/activity recognition, and human pose estimation.

Expert systems: These are highly interpretable programs designed as white-box systems. They utilize a knowledge-based approach where domain expertise provided by experts in the field is used by a knowledge engineer to populate a knowledge base, typically consisting of if-then rules. During inference, an inference engine uses this knowledge base to derive new conclusions based on observed facts.

Machine learning (ML): ML is perhaps the most transformative subfield of AI, revolutionizing the design of intelligent systems. ML algorithms can learn predictive patterns from labelled or unlabelled data autonomously, without explicit programming for specific tasks. Deep learning (DL), a subset of ML that emulates the structure and functioning of the human brain, currently represents the most promising avenue due to its effectiveness with large datasets.

Multi-agent systems: Part of distributed AI, these systems focus on interactions among autonomous entities known as agents. Agents have the capability to independently perceive their environment and can collaborate or negotiate with other agents to achieve mutually beneficial outcomes.

Natural language processing (NLP): This field employs computational techniques to comprehend, generate, and manipulate human language across various levels of linguistic analysis.

Robotics: Robotics involves the creation of physical machines that operate with varying levels of autonomy. These machines continually adapt to their surroundings through iterative processes of perception, planning, and execution.

Speech recognition: Speech recognition focuses on automated methods for processing spoken language, enhancing interactions between humans and computers.

No-code AI solutions are revolutionizing the speed of AI model development, reducing the time to mere minutes and enabling widespread integration of ML models into company workflows. These platforms cater specifically to non-technical users, allowing them to build ML models without needing to understand every step of the modelling process. While this accessibility simplifies usability, it restricts customization options. The market for no-code AI platforms, which empower individuals without specialized skills to create algorithms, is rapidly expanding. Looking forward, there will be increasing demand not just for deploying various models but potentially thousands of distinct AI applications. Users will have the capability to design and implement their own algorithms.

AI technologies represent a frontier where innovation intersects with responsibility. As these technologies evolve, best practices must continuously adapt to ensure ethical use, transparency, and accountability to leverage the trust of AI-based systems. Embracing robust data governance, rigorous testing methodologies, and ongoing monitoring are essential. Moreover, fostering interdisciplinary collaboration between technologists, ethicists, policymakers, and stakeholders will be crucial in navigating the complex ethical and societal implications of AI. By prioritizing fairness, inclusivity, and sustainability, organizations can harness the transformative potential of AI while mitigating risks and ensuring that these technologies serve the greater good. As AI continues to reshape industries and societies, adherence to these best practices will be pivotal in shaping a future where AI enhances human capabilities and advances global progress.

3.5 Tools and Best Practices for managing AI trustworthiness characteristics

Recent efforts have focused on developing methodologies across different stages of the AI lifecycle. These include designing systems with trustworthy requirements, ensuring fair and secure data collection, preprocessing, and protection, enhancing model interpretability, and implementing robust auditing and testing procedures. Based on NIST AI guidelines and the ENISA AI initiatives, the FAITH approach will emphasize human involvement in assessing trustworthiness risks. Collaborative intelligence [35], contribute towards the aim to align AI behaviour with social expectations, fostering trust and reliability in AI systems. As AI continues to evolve, ensuring its trustworthy development remains essential for its successful integration into diverse sectors while safeguarding users and society from potential harms. This section describes tools and best practices for managing AI trustworthiness characteristics.

3.5.1 Secure and resilient trustworthiness characteristic

In the life cycle of AI systems, it is critical to meet the robustness criteria for the secure and resilient systems with no errors that might exist in it. AI systems in real life might be subjected to changes in its environment, such as adversarial attacks from malicious users.

In safety-critical applications, the presence of adversarial examples poses a very dangerous situation when machine learning is being used. For instance, adversarial manipulations in an autonomous vehicle environment, like altering road signs, pose serious risks of wrong perception in the vehicle’s sensing mechanisms. For example, a slight change can lead to a system to mistake between 30 and 80 speed signs or not to identify a stop sign at all [36,37]. Further, deep learning models that apply in the identification process can be deceived by adversarial procedures. An attacker can simply misrepresent himself as an authorised user by submitting mislabelled training samples [38].

Moreover, adversarial attacks target natural language processing tasks, such as text classification, machine translation, and dialogue generation. In machine translation context, the adversarial examples are generated by paraphrasing the distinct original text, as well as the words [39]. Attackers construct a paraphrasing group of words and sentences and utilize a method to search for genuine paraphrases. Also, a gradient-guided method enhances the search in this regard. In dialogue generation, a reinforcement learning framework assists in identifying trigger inputs used to obtain deterministic output in a black box setting when the generated response needs to be semantically equivalent but different in form [40].

Also, there are cases where state-of-the-art speech-to-text systems can be adversely affected by small perturbations [41]. An attacker can successfully add a sound perturbation to a speech waveform, and these perturbations can be made by the system to say anything an attacker desires. Also, adversarial attacks can avoid the detection of YouTube’s copyright system [42]. That way, using the fingerprint created on a music piece with the help of a neural network and the features it extracts, attackers can create the perturbations, enabling the fingerprint of the song to avoid detection by the copyright holders.

These examples show that it is important to confront adversarial threats in AI systems especially within tasks that are safety sensitive. The following measures may be taken on strategies that will help improve the safety of AI systems. For instance, it is possible to find that YouTube's music copyright system is susceptible to unnoticeable noise, and this can be prevented by adversarial training. Since there is the problem of using inappropriate language when designing chatbots, it is clear that the methods of evaluating and enhancing chatbots can be made stronger. Moreover, proposing the adversarial environments for autonomous driving is the goal to increase safety.

Furthermore, to increase the level of AI security, the AI systems must be designed to be less susceptible to adversarial threats and must be able to respond to attacks that generate high confidence errors. Additionally, we need ways to detect various risks, abnormalities, and novel features in the performance of AI systems in order to guarantee they work correctly. Last but not the least is the definition of safety objectives. Monitoring and directing steering models to achieve the established safety metrics both inside and outside environments and to make sure that models meet human goods.

According to Table 10, several tools and frameworks have been developed to address adversarial threats and improve the robustness of AI models. **Adversarial Robustness Toolbox (ART)** [43] is widely used to evaluate and enhance the security of AI systems. is It is designed to implement adversarial threats, including evasion, poisoning, extraction, and inference attacks. Similarly, **Cleverhans** [44] provides implementations of adversarial attacks and supports multiple frameworks such as JAX, PyTorch, and TensorFlow 2, offering flexibility for researchers and developers. Additionally, **DeepRobust** [45] is a Python library built using PyTorch that focuses on adversarial attacks and defences for images and graphs. This library serves as a resource for implementing both attack and defence strategies in adversarial machine learning. Additionally, **RobustBench** [46] offers an evaluation platform using the Autoattack algorithm for different adversarial training models and provides well-trained robust models by various adversarial training methods. **Advbox** [47] provides adversarial attack implementations using PyTorch and TensorFlow. Similarly, **Advertorch** [48] offers PyTorch-based implementations for adversarial attacks, and **Foolbox** [49] supports adversarial attack creation across JAX, PyTorch, and TensorFlow 2. In the domain of testing and evaluation, **Giskard** [50] is an AI testing platform for detecting performance, bias, and security issues. Other notable tools include **TextAttack** [51], a Python framework for adversarial attacks and adversarial training in natural language processing, and **Torchattacks** [52], a PyTorch library for adversarial attack implementations.

Table 10: A summary of open-source tools to assess the robustness, security and resilience of AI models along with their relevant GitHub repositories

Tool	Description	GitHub
Adversarial Robustness Toolbox (ART) [43]	A library that provides implementations of adversarial threats for AI such as evasion, poisoning, extraction and inference.	https://github.com/Trusted-AI/adversarial-robustness-toolbox

Cleverhans [44]	Provides implementations of adversarial attacks in three different frameworks (JAX, PyTorch and TF2).	https://github.com/cleverhans-lab/cleverhans
DeepRobust [45]	A Python library built using the PyTorch framework that provides implementations for adversarial attacks and defenses for images and graphs.	https://github.com/DSE-MSU/DeepRobust
RobustBench [46]	A standardized benchmark for evaluating adversarial robustness of neural networks.	https://github.com/RobustBench/robustbench
Advbox [47]	A toolkit based on PyTorch and TensorFlow that provides adversarial attack implementations.	https://github.com/advboxes/AdvBox
Advertorch [48]	A toolkit based on PyTorch that provides implementations of adversarial attacks.	https://github.com/BorealisAI/advertorch
Foolbox [49]	A Python library to create adversarial attacks in three different frameworks (JAX, PyTorch and TF2).	https://github.com/bethgelab/foolbox
Giskard [50]	An open-source AI testing platform that automatically detects performance, bias and security issues in AI and LLM models.	https://github.com/Giskard-AI/giskard
TextAttack [51]	A Python framework for adversarial attacks, data augmentation and adversarial training in NLP.	https://github.com/QData/TextAttack
Torchattacks [52]	A PyTorch library that provides implementations of adversarial attacks to test the robustness of machine learning models.	https://github.com/Harry24k/adversarial-attacks-pytorch

3.5.2 Valid and Reliable - Safe trustworthiness characteristic

Validation is the "confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled" (Source: ISO 9000:2015 [53].While reliability is defined in the same standard as the "ability of an item to perform as required, without failure, for a given time interval, under given conditions" (Source: ISO/IEC TS 5723:2022 [54]).

This characteristic meets the need for AI systems to operate as desired depending on conditions that are present at a given time as well as over time and still guarantee validity of outcome. If an AI system is to be considered as reliable then, the system must guarantee consistent and repeated performances to a given set of inputs. Also, the validation processes need to ensure that the model's predictions or decisions made correspond with the goal and application of use.

Reliability is also tightly connected to safety as, in safety critical environments like healthcare, self-driving cars, or finance, the system needs to be reliable as the ever-increasing amount of tasks requires reliably correct outcomes. If AI systems are sound, they are expected to perform optimally, such that failure incidences or undesired effects are rarely observed. By testing, validation, and monitoring methods it becomes possible to enhance reliability as well as safety of AI hence making it possible to deploy the secured systems in real life situations.

Building valid and reliable AI systems necessitates addressing uncertainty at every stage of development. From data preparation to model deployment, development assumptions can influence accuracy and undermine trust. By identifying sources of uncertainty and employing mitigation strategies—categorized as data-driven, architecture-modifying, and post-hoc approaches—developers can design more robust and transparent AI systems [55]. Table 11 outlines a variety of tools and frameworks that support uncertainty estimation. For example, **TensorFlow Probability (TFP)** [56] integrates probabilistic reasoning and statistical methods into TensorFlow for uncertainty-aware modelling. **Pyro** [57], built on PyTorch, provides a deep probabilistic programming library for Bayesian inference and uncertainty estimation. **IBM's UQ360** [58] is an open-source toolkit offering explainable and interpretable uncertainty quantification methods across diverse applications. **GPyTorch** [59] employs Gaussian processes for robust uncertainty estimation in regression tasks. The **Uncertainty Toolbox** [60] delivers predictive uncertainty quantification, calibration, and visualization tools, while **NGBoost** [61], based on gradient boosting, generates probabilistic predictions with natural uncertainty estimates.

Table 11: A summary of open-source tools to assess the uncertainty of AI models along with their relevant GitHub repositories.

Tool	Description	GitHub
TensorFlow-Probability [56]	A TensorFlow library for probabilistic reasoning and uncertainty-aware modeling.	https://github.com/tensorflow/probability
Pyro [57]	A probabilistic programming library in PyTorch for Bayesian inference and uncertainty estimation.	https://github.com/pyro-ppl/pyro

UQ360 [58]	A toolkit offering explainable uncertainty quantification methods across various applications.	https://github.com/IBM/UQ360
GPyTorch [59]	A Gaussian process library in PyTorch for scalable uncertainty estimation in regression tasks.	https://github.com/cornellius-gp/gpytorch
Uncertainty-Toolbox [60]	A Python library for predictive uncertainty quantification, calibration, and visualization.	https://github.com/uncertainty-toolbox/uncertainty-toolbox
NGBoost [61]	A gradient-boosting framework for probabilistic predictions with natural uncertainty estimates.	https://github.com/stanfordmlgroup/ngboost

3.5.3 Explainable and interpretable

Being able to justify and explain AI results is the key to enhancing the trustworthiness in AI. Recommendation systems like Amazon support the decision making of customers by recommending certain products [62]. There is a need to explain why such a constitution should be included in these systems and this process creates an enhanced level of trust and confidence. For example, RuleRec [63] is based on a joint learning procedure to provide accurate and reliable and explainable recommendations by extracting encoded rules associated with item connections such as 'Also viewed' and 'Buy together'. These explanations can make users change their behaviour and increase confidence in the system.

Natural Language Processing (NLP) is one of the categories of AI aimed at helping computers process natural languages. It encompasses a form of applications for instance dialog systems, computer or machine translation and sentiment analysis. Recent innovations in deep learning compromised the explanation of models and enhanced the correctness of a plethora of NLP tasks. The use of NLP systems may be a problem because users cannot totally trust them due to the lack of explainability and interpretability. To deal with this challenge, researchers have designed methods such as LIME that perturb input data to explain predictions in text classification models. Another approach is known as CAML that uses construction of attention mechanisms to accomplish the identification of important segments in clinical text for medical code implementation. Consumers must understand why and how a certain NLP model produces the result it does, and these methods seek to make this possible.

Table 12 provides an indicative list of several tools for explainability analysis of AI models. **BertViz** [64] specializes in visualizing attention mechanisms in Transformer-based models, offering model-specific insights into BERT outputs and attention scores. **Captum** [65], designed for PyTorch models, includes general-purpose explainability techniques such as integrated gradients, saliency maps, and smoothgrad, focusing on model-agnostic imaging applications. Similarly, **CNN-explainer** [66] is an interactive visualization system that simplifies understanding of convolutional neural networks (CNNs) for non-experts, while the **Deep-visualization-toolbox** [67] provides interactive exploration and analysis of CNNs. **Grad-CAM** [68] highlights important image regions using gradients from the final convolutional layer, making it a widely used model-agnostic imaging technique. For tabular data explainability, **InterpretML** [69] offers state-of-the-art methods for model-agnostic interpretability, while **LIME** [70] explains predictions for diverse data types, including tabular, imaging, and text, by learning interpretable models locally around predictions. **Netron** [71] acts as a model-specific viewer for visualizing neural networks and machine learning models, offering detailed structural insights. **PyTorch-CNN-Visualizations** [72] implements visualization techniques specifically for CNNs, enhancing understanding of imaging data. **SHAP** [73] uses game theory principles to provide model-agnostic explanations for tabular, imaging, and text data, offering a versatile approach to interpreting machine learning models. **IBM's AIX360** [74] is another open-source toolkit that provides a suite of explainability algorithms for AI models, including both model-specific and model-agnostic methods. Moreover, **DIG** [75] incorporates a Python toolkit for explaining graph deep learning models, focusing on graph-based data and their correspondingly complex and specialized approaches. **DeepExplain** [76] directs towards gradient-based methods and perturbation-based methods for explainability and encompasses support for various methods types.

Table 12: A summary of open-source tools and methods for explainability analysis of AI models along with their relevant GitHub repositories.

Tool	Description	Types of Data	Type of Explainability Method	GitHub
BertViz [64]	A tool that is developed to visualize the attention mechanisms of Transformer-based models.	BERT model outputs, attention scores.	Model-Specific	https://github.com/jessevig/bertviz
Captum [65]	It contains general-purpose implementations of integrated gradients, saliency maps, smoothgrad and others for PyTorch models.	Imaging	Model-Agnostic	https://github.com/pytorch/captum

CNN-explainer [66]	An interactive visualization system that was developed to assist non-experts learn about the inner workings of convolutional neural networks (CNNs).	Imaging	Visualization	https://github.com/poloclub/cnn-explainer
Deep-visualization-toolbox [67]	Provides interactive visualizations to explore and analyze convolutional neural networks (CNNs).	Imaging	Visualization	https://github.com/yosinski/deep-visualization-toolbox
Grad-CAM [68]	This explainability technique uses the gradients of the target class from the final convolutional layer to highlight the most important regions in the image.	Imaging	Model-Agnostic	https://github.com/ramprs/grad-cam
InterpretML [69]	It provides state-of-the-art machine learning explainability and interpretability techniques.	Tabular	Model-Agnostic	https://github.com/interpretml/interpret
LIME [70]	A technique designed to explain the predictions of any machine learning classifier by learning an interpretable model locally around the prediction.	Tabular, Imaging, Text	Model-Agnostic	https://github.com/marcotcr/lime-experiments

Netron [71]	A viewer for neural networks, deep learning, and machine learning models.	Neural Network Models	Model-Specific	https://github.com/lutzroeder/netron
PyTorch-CNN-Visualizations [72]	A PyTorch implementation of visualization techniques for convolutional neural networks.	Imaging	Model-Specific	https://github.com/utkuozbulak/pytorch-cnn-visualizations
SHAP [73]	An explainability technique that applies definitions from game theory to explain machine learning models.	Tabular, Imaging, Text	Model-Agnostic	https://github.com/shap/shap
AIX360 [74]	A suite of explainability algorithms for machine learning models, including both model-specific and model-agnostic methods	Tabular, Imaging, Text	Model-Specific, Model-Agnostic	https://github.com/Trusted-AI/AIX360
DIG [75]	A library for graph deep learning research, providing a unified testbed for tasks like graph generation, self-supervised learning, explainability, 3D graphs, and graph out-of-distribution.	Graph-structured data	Model-Specific, Visualization	https://github.com/divelab/DIG

DeepExplain [76]	A unified framework of perturbation and gradient-based attribution methods for Deep Neural Networks interpretability. DeepExplain also includes support for Shapley Values sampling	Imaging,Text	Model-Specific, Model-Agnostic,Visualization	https://github.com/marcoancona/DeepExplain

3.5.4 Privacy-enhanced

Privacy-preserving techniques have become popular in safeguarding sensitive information within operational systems. These methods are crucial for ensuring the confidentiality of personal data across various applications. The privacy concern surrounding personal healthcare information restricts the full exploitation of diverse healthcare data for predictive models. Federated learning holds the key regarding how data can be used across populations while not broadcasting the information [77]. Differential privacy is widely used to protect the privacy of healthcare data as well [78]. Generally, in recommender systems, the interaction of users with products might reveal sensitive information such as age and gender through rating information [79]. To overcome this, privacy preserving recommender systems based on differential privacy techniques have been proposed for preserving user privacy while generating recommendation [80].

Furthermore, research [43] has shown many privacy issues connected with the large language model like GPT-2, exposing that there are few attacks to extracting private information. Consequently, differential privacy techniques, and specifically DP-SGD have been applied more often for training large-scale language models [44]. These methods are designed to reduce privacy risk while keeping model accuracy and training speed.

Several tools support privacy, security, and distributed learning for diverse types of data (Table 13). **Diffprivlib** [81] provides differential privacy techniques for machine learning models, specifically supporting numerical tabular data. **FATE** [82] is an industrial-grade federated learning framework designed for collaborative learning on tabular and imaging data while preserving privacy. Similarly, **FedML** [83] and **Flower** [84] enable federated learning, distributed training, and model serving across numerical tabular, imaging, and text data. **Helib** [85], on the other hand, focuses on implementing homomorphic encryption for efficient evaluation on numerical tabular data. For privacy-preserving training **PySyft** [86] enables secure machine learning and statistical analysis for tabular, imaging, and text data. Other tools like **Open Policy Agent** [87] act as policy engines, ensuring secure data access based on user permissions. **TensorFlow Federated** [88] provides an open-source framework for decentralized computations, supporting numerical tabular, imaging, and text data. **Tensorflow Privacy** [89] is a Python library that includes implementations of TensorFlow optimizers for training machine learning models with differential privacy.

Table 13: A summary of open-source tools to assess the privacy of AI models along with the relevant GitHub repositories.

Tool	Description	Types of Supported Data	GitHub
Diffprivlib [81]	An open-source general-purpose library developed by IBM that provides implementations of differential privacy techniques for machine learning models.	Tabular	https://github.com/IBM/differential-privacy-library
FATE [82]	An industrial-grade open-source federated learning framework that enables enterprises and institutions to collaborate on data while protecting data security and privacy.	Tabular, Imaging	https://github.com/FederatedAI/FATE
FedML [83]	A research-oriented library that provides large-scale distributed training, model serving, and federated learning.	Tabular, Imaging, Text	https://github.com/FedML-AI/FedML
Flower [84]	It is a framework that is designed to simplify the development of federated learning systems.	Tabular, Imaging, Text	https://github.com/adap/flower
Helib [85]	An open-source software library that implements homomorphic encryption. It also includes optimizations for efficient homomorphic evaluation.	Tabular	https://github.com/homenc/HElib

Open Policy Agent [86]	It is an open-source general-purpose policy engine that enables only authorized users to access certain data.	Access Logs, Policy Data, and User Data	https://github.com/open-policy-agent/opa
PySyft [87]	A library that enables users to perform any statistical analysis for machine learning by using non-public information.	Tabular, Imaging, Text	https://github.com/OpenMined/PySyft
Tensorflow Federated [88]	An open-source framework used in machine learning and other computations on decentralized data.	Tabular, Imaging, Text	https://github.com/google-parfait/tensorflow-federated
Tensorflow Privacy [89]	Library for training machine learning models with privacy for training data	Tabular, Text, Imaging	https://github.com/tensorflow/privacy
DPA [90]	A Python-based toolkit with a set of metrics to assess data privacy risks	Tabular	https://github.com/FAITH-FORTH/DPA.

The **Data Privacy Assessment (DPA) toolkit** [90] was developed by FORTH to help users assess data privacy risks and the utility of real and anonymized (or synthetic) data. This toolkit is particularly useful for organizations and researchers seeking to assess how effectively anonymization techniques preserve privacy while maintaining data utility. DPA offers privacy risk metrics which aim to assess how vulnerable a dataset is to privacy breaches, after anonymization, by assessing a variety of key metrics.

3.5.5 Accountable and transparent

Accountability in AI refers to the trustworthiness of AI technologies and the assignment of responsibility when these technologies fail to meet expectations. The process of auditing creates an important measure of accountability as the AI systems themselves are evaluated on a number of factors. As derived from the IEEE standard for software development, audits mean independent examination of software products and processes in relation to the rule and regulation and standard. Greater third-party independence in external audits [91] provide preferable external views, but can be restricted by poor access to key internal information such as model training data, presenting relative difficulties in effective assessment and potential post-deployment delays of issue identification. The internal auditors, those who belong to the organization that develops and uses the AI system, can have access to vast data in the company that can be checked comprehensively before implementation and serve as the primary source of information for decision-makers. However, internal audits may face bias concerns due to shared interests with the audited party.

Transparency involves detailed documentation of an AI system's entire lifecycle, along with the underlying tasks and procedures that define its operation. Ensuring transparency from the beginning of the development of an AI system is crucial to eliminate any uncertainty about its functionality and application. In this respect, it is strictly interconnected with the traceability of data as well as the traceability of AI systems. Finally, Transparency also ensures that an AI system is designed to be reproducible and auditable, laying the groundwork for accountability and responsibility. Key Aspects of Traceability: Data Lineage: documenting the source, preprocessing, and transformations applied to datasets; Model Lineage: capturing metadata like model architecture, training data, training duration, and hyperparameter; version Control: ensuring every change to data, code, and model parameters is tracked; decision Path Tracking: logging intermediate outputs of the model for debugging and understanding predictions. The following table reports open-source tools that support AI developers in leveraging transparency of their AI model development pipeline.

The listed open-source tools (Table 14) collectively support the traceability characteristic of AI trustworthiness by promoting transparency, accountability, and detailed documentation across various aspects of AI systems. **Model Cards** [92] provide structured documentation detailing the performance, strengths, and limitations of machine learning models, making them essential for stakeholders to assess a model's suitability for their needs. Similarly, **Datasheets for Datasets** [93] ensure dataset transparency by offering comprehensive details about their motivation, composition, collection processes, and limitations, addressing critical ethical considerations and enabling responsible dataset usage. **FactSheets** [94] expand this concept by thoroughly documenting the functionalities, intended use, performance, safety, and security of AI systems, while also covering their lifecycle, including training, deployment, and testing. This makes them a robust tool for organizations focused on governance and compliance in regulated industries. The **Model Card Toolkit** [95] complements these by simplifying the creation of Model Cards, integrating seamlessly into workflows, and making the framework accessible to a wide range of teams. The **FAITH AI Model Passport** adds an innovative layer of automation, encapsulating metadata crucial for model identification, validation, and traceability, thereby enhancing transparency and reproducibility. It also supports the integration with popular frameworks, ensuring end-to-end traceability across AI pipelines.

Table 14: List of open-source tools supporting the traceability AI trustworthiness characteristic.

Tool	Description	GitHub
Model Cards [92]	Provides detailed documentation for machine learning models, including details on the performance characteristics, strengths and limitations.	https://github.com/tensorflow/model-card-toolkit

Datasheets for Datasets [93]	Provides comprehensive information about the datasets used during the training process of the machine learning models, describing their motivation, composition, collection process, recommended use cases and other relevant information.	https://github.com/fau-masters-collected-works-cgarbin/datasheet-for-dataset-template?tab=readme-ov-file
FactSheets [94]	Contains important information about all relevant functionalities of an AI-based system, such as its intended use, performance, safety and security levels. Moreover, it documents how the AI service was created, trained and deployed, along with the testing procedure that is used, how it might respond to untested scenarios and any other ethical concern for its usage.	https://github.com/IBM/ai-governance-factsheet-samples
Model Card Toolkit [95]	An open-source implementation of Model Cards.	https://github.com/tensorflow/model-card-toolkit
FAITH AI model passport	It encapsulates in an automatic way vital metadata crucial for model identification and validation and enhances transparency and enables reproducibility. It is a pythonic library supported by well-known frameworks such as MLFlow and DVC that guarantees end-to-end traceability.	https://github.com/FAITH-FORTH/AIPassport

3.5.6 Human oversight

The comprehensive understanding of human behaviors, preferences, and expectations in the development of AI tools is an important element of the FAITH framework. To achieve this, it is crucial to implement effective human oversight instruments and methodologies that allow for the collection of both quantitative and qualitative data from users. This data informs the design and functionality of AI systems, ensuring they are user-centric and aligned with societal norms and values. Key methodologies in this context include co-production workshops and the use of surveys and questionnaires, each offering unique advantages for gathering insights from human participants.

Co-production workshops are interactive sessions where relevant stakeholders, including end-users, developers, and subject matter experts, collaboratively participate in the AI design and development process. These workshops facilitate a deeper understanding of user needs and provide a platform for direct feedback and iterative design. Through co-production, stakeholders can influence the trajectory of any development, ensuring the technology is not only technically robust but also socially and ethically aligned. It is widely accepted that co-production is a powerful approach for integrating diverse perspectives, fostering innovation, and building consensus around the ethical implications of AI tools.

On the other hand, surveys and questionnaires are essential tools for collecting quantitative data on user preferences, behaviours, and expectations. These instruments are particularly useful for reaching a broad audience, allowing developers to gather statistically significant data that can be generalized across larger populations. Well-designed surveys can uncover patterns in user expectations and highlight potential biases in AI systems, providing a foundation for refining algorithms and user interfaces. Surveys and questionnaires are crucial for understanding user demographics and personality traits, which are key to designing AI tools that are inclusive and representative of diverse user groups.

3.5.7 Fair - with harmful bias managed

The relation between risk estimation for technical threats of trustworthiness and fairness Risk estimation for technical threats of trustworthiness involves the identification, assessment, and mitigation of risks that can consequence the technical reliability and overall trustworthiness of AI systems. This process addresses various types of risks, including algorithmic biases that can result in unfair treatment of certain groups, data quality issues stemming from inaccuracies or biases in the training and testing datasets, and model robustness challenges related to the AI system's ability to perform consistently across diverse scenarios. Additionally, security threats that could exploit vulnerabilities to compromise the AI system and issues related to transparency and explainability, which ensure that AI model decisions are understandable and justifiable, are also critical components of this risk estimation process.

Analyses of existing models show that machine learning models have various biases in the different fields. For example, face recognition systems have been found to perform better with white faces than the darker ones and are biased with the gender as well [29][30]. Additionally, algorithmic biases are common in natural language processing tasks. Similar patterns are found in sentence embeddings and co-reference resolution systems, which show a higher accuracy for gendered pronouns linked to pro-stereotypical entities. Language models learn gender discrimination from text data, generating words that reflect gender stereotypes differently for males and females. Also, voice recognition systems show gender bias, processing male voices more accurately than female voices, affecting applications in medical voice recognition [96] and vehicle voice control systems [97].

The Fairness characteristic, as outlined in the NIST AI Risk Management Framework (RMF) [98], highlights the importance of managing harmful biases to ensure that AI systems treat all individuals equitably. This characteristic emphasizes the need for thorough bias identification to recognize biases in AI models and their training data, and bias mitigation strategies to reduce or eliminate these biases. Ensuring equity involves making sure that AI systems do not disproportionately consequence any specific group negatively. Accountability is another key aspect, holding developers and deployers of AI systems responsible for ensuring fairness.

Continuous monitoring is essential to regularly check and update AI systems, guaranteeing ongoing fairness throughout their lifecycle.

3.5.7.1 Categories of bias

According the **NIST Standard for Identifying and Managing Bias in Artificial Intelligence** [99], there are three main categories of statistical and computational bias, which are related to: (i) Processing/Validation, (ii) Use and interpretation, and (iii) Selection and sampling. These categories are described next along with the related biases.

- **Processing/Validation**: Arises during the stages of data processing and validation, where biases can be introduced or amplified in the data handling or algorithmic processes. This category of bias can skew the distribution of prediction outputs compared to the actual distribution of the prediction targets. Types of Processing/Validation biases are described next. Amplification bias arises when there is a discrepancy between the distribution of prediction outputs and the prior distribution of the prediction target. This skewed distribution introduces a measurable bias, highlighting the need for robust data processing and validation techniques to mitigate its consequence. Similarly, Inherited bias occurs when machine learning applications generate biased outputs that serve as inputs for other algorithms, propagating the initial bias. The Model Selection bias is introduced during the selection of a single "best" model from a large set of models using various predictor variables. This bias also manifests when an explanatory variable has a weak relationship with the response variable, leading to skewed model performance.
- **Use and Interpretation**: Occurs when biases emerge from the way data is utilized and interpreted within AI systems. This can include selection biases based on activity levels or concept drift when systems are used outside their intended domains. Examples of use and interpretation types of biases are described next. The Survivorship bias refers to the tendency to focus on items or individuals that "survive" a selection process, overlooking those that did not. Activity bias, a type of selection bias, occurs when systems derive their training data from the most active users, neglecting less active or inactive users. Concept Drift involves the use of a system outside its planned domain of application, leading to performance discrepancies between laboratory settings and real-world environments. Emergent bias arises naturally during system use. Content Production bias stems from structural, lexical, semantic, and syntactic differences in user-generated content.
- **Selection and sampling**: Arises during the selection of data samples or groups for training and testing, leading to unrepresentative datasets. This can cause the model to perform poorly on underrepresented groups or scenarios. Examples of selection and sampling types of biases are described next. The Data Generation bias occurs from the addition of synthetic or redundant data samples to a dataset, leading to potential skewing of results. Detection bias involves the existence of systematic differences in the outcome determination between groups, causing over- or underestimation of effects. Ecological Fallacy arises when inferences about individuals are made based on group membership, leading to potentially inaccurate conclusions. Evaluation bias

arises when testing populations do not equally represent the user population or when inappropriate performance metrics are used. The Exclusion bias occurs when specific user groups are excluded from testing and analyses, leading to unrepresentative results. Measurement bias arises when features and labels are proxies for desired quantities, potentially introducing group-specific noise. Popularity bias, a form of selection bias, occurs when more popular items are overexposed while less popular items are under-represented. Representation bias arises from non-random sampling of subgroups, causing trends estimated for one population to not be generalizable to a new population. Simpson's Paradox, a statistical phenomenon, occurs when the association between two variables changes when controlled for another variable. Temporal bias arises from changes in populations and behaviours over time, leading to outdated or skewed results. Uncertainty bias occurs when algorithms favour well-represented groups in the training data due to less prediction uncertainty.

ISO24027 [100] provides a framework to address various data-related challenges and biases in AI systems, ensuring the development of robust and reliable models. It emphasizes the importance of handling missing features and labels, implementing effective data aggregation, and addressing non-representative sampling to maintain the quality and representativeness of datasets. The standard highlights the role of distributed training, focusing on decentralized data processing and privacy preservation. It also addresses data biases, including selection bias, sampling bias, non-response bias, and coverage bias, while recognising additional sources of bias such as confounding variables and Simpson's paradox. Proper data processing, including cleaning, normalisation, and label accuracy, is critical for mitigating biases and ensuring the validity of AI models. By incorporating these principles and tools, ISO 24027 promotes the development of transparent, fair, and effective AI systems that are resilient to data-related uncertainties and biases. It covers the following categories of bias:

- **Selection bias**: Bias introduced due to non-random sampling of data, leading to unrepresentative datasets.
- **Sampling bias**: Arises when certain subsets of the population are over- or underrepresented in the dataset.
- **Coverage bias**: Occurs when specific groups or populations are systematically excluded or underrepresented in the dataset.
- **Non-response bias**: Results from the failure of certain individuals or units to respond during data collection, causing incomplete datasets.
- **Confounding variables**: Variables that influence both the dependent and independent variables, potentially distorting outcomes.
- **Simpson's paradox**: Situations where trends observed in aggregated data are reversed when analysed at a disaggregated level, leading to misleading conclusions.
- **Data bias**: General biases embedded in the data collection, representation, or processing pipeline, which can result in skewed model outcomes.
- **Other sources of bias**: Includes factors such as cultural, temporal, or technical artifacts, and environmental influences that can distort datasets.

3.5.7.2 Open-source tools and methods for data and AI model bias assessment

The landscape of open-source libraries dedicated to addressing bias in machine learning is rich and varied, comprising tools designed for bias detection, mitigation, and fairness enhancement across various data types, primarily tabular, with some extending to imaging.

In Table 15, a variety of open-source tools that have been developed to assess and mitigate bias in data and AI models are presented. The **AI Fairness 360 (AIF360)** [101], developed by IBM, is one of the most widely used tools for detecting and mitigating bias in AI models. It provides a collection of metrics and algorithms for bias assessment and mitigation. Similarly, **Fairlearn** [102] is a Python package that offers AI developers a rich collection of metrics and algorithms that can be used for the assessment of the model's fairness. Both tools can assist data scientists in addressing all the potential fairness concerns that arise either in their datasets or to their AI models. **Themis-ML** [103] is a Python library that is designed to detect and mitigate biases in tabular machine learning models. Additionally, **BiasOnDemand** [104] enables researchers and AI developers with the generation of synthetic datasets with different types of biases. **Aequitas** [105] audits machine learning models for fairness by analysing predictions, identifying biases, and providing fairness metrics and visualizations for different demographic groups. **Fairness Measures** [106] evaluates model outcomes using metrics like demographic parity, equalized odds, and predictive parity to ensure unbiased and equitable predictions across groups. **FAT Forensics** [107] is a Python toolbox that is built upon SciPy and NumPy and is used for the fairness, accountability, and transparency evaluation of predictive AI models. **REVISE** [108] provides tools to measure and mitigate biases in imaging datasets, addressing potential fairness and bias concerns in computer vision applications.

Table 15: A summary of open-source tools to assess the bias in data and AI models along with their relevant GitHub repositories.

Tool	Description	GitHub
AIF360 [101]	An open-source toolkit developed by IBM that provides techniques for the detection and mitigation of bias that is presented in machine learning models.	https://github.com/Trusted-AI/AIF360
Fairlearn [102]	An open-source toolkit that is used for the assessment and improvement of fairness in machine learning models.	https://github.com/fairlearn/fairlearn

Themis-ML [103]	A Python library that has developed fairness-aware machine learning algorithms for the detection and mitigation of biases that are presented in machine learning models.	https://github.com/cosmicBboy/themis-ml
BiasOnDemand [104]	A Python package that generates synthetic datasets with different types of bias.	https://github.com/rcrupiISP/BiasOnDemand
Aequitas [105]	Provides an easy-to-use and transparent tool for auditing predictors of ML models, as well as experimenting with "correcting biased models" using Fail ML methods in binary classification settings.	https://github.com/dssg/aequitas
Fairness Measures [106]	An open-source Python toolkit that provides datasets and software for detecting algorithmic discrimination.	https://github.com/megantosh/fairness_measures_code/tree/master
FAT Forensics [107]	A modular Python toolbox for algorithmic fairness, accountability and transparency.	https://github.com/fat-forensics/fat-forensics
REVISE [108]	A tool for measuring and mitigating bias in visual datasets.	https://github.com/princetonvisualai/revise-tool
DBDM (Data Bias Detection and Mitigation) [109]	A Python-based toolkit for detecting and mitigating pre-training data bias	https://github.com/FAITH-FORTH/DBDM
MANDALA (Measure post-trAiniNg Data And modeL biAs) [110]	A Python-based toolkit for detecting post-training data and model bias	https://github.com/FAITH-FORTH/MANDALA

The **DBDM (Data Bias Detection and Mitigation) toolkit** [109], developed by FORTH, focuses on the detection and mitigation of pre-training data bias. It employs a suite of 15 statistical-based metrics to offer a holistic view of data bias before the AI model training process. It also supports cluster analysis to identify and analyze biases within clusters. As far as data bias mitigation is concerned, the toolkit utilizes synthetic data to populate the underrepresented groups. In addition, **MANDALA (Measure post-trAiniNg Data And modeL biAs)** [110] is another tool developed by FORTH which aims to detect post-training data and model biases through a set of 13 implemented metrics. Both tools will be linked with the **FAITH AI model passport**.

3.6. Metrics adopted in trustworthiness management tools

The tools mentioned in the section 3.5, use a variety of measurements, factors, and scales to evaluate the different aspects of the reliability of AI systems. In this section, we will describe the metrics that facilitate the evaluation of various aspects of trust in AI systems. In this way, a comprehensive picture of how these tools operate and support the development of best practices for AI will be provided.

3.6.1 Metrics for data bias detection

A variety of statistical-based metrics has been proposed in the literature for data bias detection (Table 16) [111][112]. Examples of such metrics, include: (i) the **Class Imbalance (CI)** which evaluates the imbalance between the groups within a facet, (ii) the **Difference in Proportions of Labels (DPL)** which measures the disparity in positive outcomes between the groups in the facet, (iii) the **Demographic Disparity (DD)** which computes the disparity for specific groups in the facet, (iv) the **Conditional Demographic Disparity (CDD)** which examines demographic disparities within subgroups, (v) the **Kullback-Leibler (KL) divergence** which estimates the divergence between probability distributions of facets and outcomes, (vi) the **Jensen-Shannon (JS) Divergence** which is similar to KL but a symmetrized version, (vii) the **Total Variation Distance (TVD)** which measures the distance between distributions of facets and outcomes, (viii) the **Kolmogorov-Smirnov (KS)** metric which assesses the statistical distance between distributions, (ix) the **Normalized Mutual Information (NMI)** and the **Normalized Conditional Mutual Information (NCMI)** which measure the information shared between categorical variables, normalized over possible outcomes, (x) the **Binary Ratio (BR)** and the **Binary Difference (BD)** which calculate ratios and differences in positive outcomes between binary groups, (xi) the **Conditional Binary Difference (CBD)** which analyses disparities within subgroups, (xii) the **Pearson Correlation (CORR)** which determines the linear correlation between the facet and the outcome, and (xiii) the **Logistic Regression (LR) coefficient** which assesses the influence of the facet to the outcome through a logistic regression model.

Table 16: The list of metrics for data bias detection.

No	Metric (acronym)	Short description	Reference
1	Class Imbalance (CI)	Evaluates the imbalance between the groups in the facet	https://link.springer.com/chapter/10.1007/978-3-642-23166-7_12
2	Difference in Proportions of Labels (DPL)	Measures the disparity in the positive outcomes between the groups in the facet	https://books.google.gr/books/about/Applied_Regression_Analysis_and_Other_Mu.html?id=v590AgAAQBAJ&redir_esc=y
3	Demographic Disparity (DD)	Computes the disparity for a specific group	https://arxiv.org/abs/1412.3756

4	Conditional Demographic Disparity (CDD)	Examines demographic disparities within subgroups	https://fairmlbook.org /
5	Kullback-Leibler divergence	Estimates the Kullback-Leibler (KL) divergence between the probability distributions of the facet and the outcome	https://www.jstor.org/stable/2236703
6	Jensen-Shannon (JS) divergence	Estimates the Jensen-Shannon (JS) divergence between the probability distributions of the facet and the outcome	https://ieeexplore.ieee.org/document/61115
7	Total Variation Distance (TVD)	Measures the distance between the probability distributions of the facet and the outcome	https://ecommons.cornell.edu/items/88a62f81-14bf-443a-9c35-9bf85b32bcab
8	Kolmogorov-Smirnov (KS) metric	Assesses the statistical distance between the probability distributions of the facet and the outcome	https://www.tandfonline.com/doi/abs/10.1080/01621459.1951.10500769
9	Normalized Mutual Information (NMI)	Measures the information shared between two categorical variables, normalized to a range of [0, 1] where 1 indicates perfect correlation and 0 indicates no correlation	https://www.jmlr.org/papers/volume3/strehl02a/strehl02a.pdf
10	Normalized Conditional Mutual Information (NCMI)	Measures the mutual information between two categorical variables, conditioned on a third, normalized over the possible outcomes of the conditioning variable	https://jmlr.csail.mit.edu/papers/volume11/vinh10a/vinh10a.pdf
11	Binary Ratio (BR)	Computes the ratio of positive outcomes between two binary groups	https://www.science.org/doi/10.1126/science.187.4175.398
12	Binary Difference (BD)	Calculates the difference in proportions of positive outcomes between two binary groups to detect disparities	https://ieeexplore.ieee.org/document/4909197
13	Conditional Binary Difference (CBD)	Computes the binary difference, conditioned on another categorical feature, to analyze disparities within subgroups	https://arxiv.org/abs/1610.02413
14	Pearson Correlation (CORR)	Determines the linear correlation between two ordinal features, with values ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation)	https://link.springer.com/chapter/10.1007/978-3-642-00296-0_5
15	Logistic Regression (LR) coefficient	Fits a logistic regression model to predict a multi-labeled outcome from a binary protected feature to assess the influence of the feature on the outcome	https://onlinelibrary.wiley.com/doi/book/10.1002/9781118548387

3.6.2 Metrics for AI model bias detection

A diverse set of statistical-based metrics has been proposed in the literature for the detection of AI model bias [113][114]. Examples of such metrics, include: (i) the **Difference in Positive**

Proportions in Predicted Labels (DPPL) which measures the difference in the proportion of positive predictions between the favoured facet a and the disfavoured facet d, (ii) the **Disparate Consequence (DI)** which measures the ratio of proportions of the predicted labels for the favoured facet, say a and the disfavoured facet, say d, (iii) the **Conditional Demographic Disparity in Predicted Labels (CDDPL)** which measures the disparity of predicted labels between the facets as a whole, but also by subgroups, (iv) the **Counterfactual Fliptest (FT)** which examines each member of facet d and assesses whether similar members of facet a have different model predictions, (v) the **Accuracy Difference (AD)** which measures the difference between the prediction accuracy for the favoured and disfavoured facets, (vi) the **Recall Difference (RD)** which compares the recall of the model for the favoured and disfavored facets, (vii) the **Difference in Conditional Acceptance (DCAcc)** Compares the observed labels to the labels predicted by a model. Assesses whether this is the same across facets for predicted positive outcomes (acceptances), (viii) the **Difference in Acceptance Rates (DAR)** which measures the difference in the ratios of the observed positive outcomes (TP) to the predicted positives (TP + FP) between the favoured and disfavoured facets, (ix) the **Specificity difference (SD)** which compares the specificity of the model between favoured and disfavored facets, (x) the **Difference in Conditional Rejection (DCR)** which compares the observed labels to the labels predicted by a model and assesses whether this is the same across facets for negative outcomes (rejections), (xi) the **Difference in Rejection Rates (DRR)** which measures the difference in the ratios of the observed negative outcomes (TN) to the predicted negatives (TN + FN) between the disfavoured and favoured facets, (xii) the **Treatment Equality (TE)** which measures the difference in the ratio of false positives to false negatives between the favoured and disfavoured facets, and (xiii) the **Generalized entropy (GE)** which measures the inequality in benefits b assigned to each input by the model predictions.

3.6.3 Metrics for AI model performance assessment

Table 17 summarizes the list of performance measures which has been identified by the OECD [115] (along with some additional propositions), where each performance measure is mapped with a learning type (i.e. supervised, unsupervised, reinforcement, all). For general performance evaluation, examples of commonly used metrics include the **accuracy** [116], **F-Score** [116], **precision** [116], **recall** [116], and **ROC-AUC** [117], which are widely applied across supervised and semi-supervised learning tasks to evaluate classification models on various datasets. For regression tasks, metrics such as the **Mean Squared Error (MSE)** [118] and the **Root Mean Squared Error (RMSE)** [118] are standard for quantifying deviations between predicted and actual labels. In specialized domains like image segmentation, metrics such as the **Dice Score** [119], the **Intersection over Union (IoU)** [120], and the **Hausdorff Distance** [121] are critical for measuring the overlap between predicted and ground truth segmentation masks. In natural language processing (NLP), metrics like **BLEU** [122], **ROUGE** [123], and **BERTScore** [124] are used to evaluate tasks such as machine translation and text summarization. Reliability and generalization are often assessed using metrics like the **Expected Calibration Error (ECE)** [125], which evaluates the alignment of predicted probabilities with ground truth, and the **Out-of-Distribution (OOD) generalization** [126], which measures model performance on previously unseen data.

Table 17: List of Metrics for AI Model Performance Assessment.

No	Title	Data type	Learning type	Website
1	Accuracy	all	Supervised	https://en.wikipedia.org/wiki/Accuracy_and_precision
2	Hausdorff Distance	imaging	Supervised, Unsupervised	https://en.wikipedia.org/wiki/Hausdorff_distance
3	Average Surface Distance	imaging	Supervised, Unsupervised	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5817231/
4	Mean Intersection over Union (IoU)	imaging	Supervised	https://giou.stanford.edu/
5	Receiver Operating Characteristic Curve (ROC) and Area Under the Curve (AUC)	tabular	Supervised	https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
6	Bilingual Evaluation Understudy (BLEU)	text	Supervised	https://dl.acm.org/doi/10.3115/1073083.1073135
7	Precision	tabular	Supervised	https://en.wikipedia.org/wiki/Precision_and_recall
8	Recall-Oriented Understudy for Gisting Evaluation (ROUGE)	text	Supervised	https://aclanthology.org/W04-1013.pd
9	Recall	tabular	Supervised	https://en.wikipedia.org/wiki/Precision_and_recall
10	Mahalanobis Distance	tabular	Supervised	http://library.isical.ac.in:8080/xmlui/bitstream/handle/10263/6765/Vol02_1936_1_Art05-pcm.pdf
11	Anonymity Set Size	tabular	Supervised	https://link.springer.com/article/10.1007/BF00206326
12	Equal performance	tabular	All	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6594166/
13	Time until Adversary's Success	time-series	Reinforcement	https://www.semanticscholar.org/paper/An-Analysis-of-the-Degradation-of-Anonymous-Wright-Adler/b20053ebb8ef3623abe6020251e84e36c6920181
14	Amount of Leaked Information	time-series	Unsupervised	https://arxiv.org/pdf/1512.00327.pdf
15	Word Error Rate (WER)	text	Supervised	https://www.sciencedirect.com/science/article/abs/pii/S0167639301000413?via%3Dihub
16	Consensus-based Image Description Evaluation (CIDEr)	text	Supervised	https://arxiv.org/abs/1411.5726
17	F-score	tabular	Supervised	https://en.wikipedia.org/wiki/F-score
18	SacreBLEU	text	Supervised	https://github.com/mjpost/sacreBLEU

19	Perplexity	text	Unsupervised	https://towardsdatascience.com/perplexity-in-language-models-87a196019a94?gi=1b1578ee91a1
20	Exact Match	text	Supervised	https://huggingface.co/spaces/evaluate-metric/exact_match#:~:text=Metric%3A%20exact_match,JSON%2Dformatted%20list%20as%20input
21	Adjusted Rand Index (ARI)	tabular	Unsupervised	https://econpapers.repec.org/article/sprjclass/v_3a2_3ay_3a1985_3ai_3a1_3ap_3a193-218.htm
22	Mean Per Joint Position Error (MPJPE)	imaging	Supervised	https://courses.grainger.illinois.edu/ece445zjui/getfile.asp?id=19050
23	Sparsity	Imaging	Unsupervised	https://arxiv.org/pdf/2210.03683.pdf
24	Equality of Opportunity Difference (EOD)	tabular	All	https://arxiv.org/abs/1610.02413
25	Conditional Entropy	time-series	Unsupervised	https://dl.acm.org/doi/10.1145/1314333.1314347
26	Stability	tabular	All	https://arxiv.org/abs/2108.13624
27	Out-of-distribution (OOD) generalization	tabular	Supervised	https://arxiv.org/abs/2108.13624
28	Cross-lingual Natural Language Inference (XNLI)	text	Supervised	https://aclanthology.org/D18-1269.pdf
29	Translation Edit Rate (TER)	text	Supervised	https://www.cs.umd.edu/~snover/pub/amta06/ter_amta.pdf
30	Metric for Evaluation of Translation with Explicit ORdering (METEOR)	text	Supervised	https://www.researchgate.net/publication/228346240_METEOR_An_automatic_metric_for_MT_evaluation_with_high_levels_of_correlation_with_human_judgments
31	Mean Squared Error (MSE)	tabular	Supervised	https://en.wikipedia.org/wiki/Mean_squared_error
32	Crosslingual Optimized Metric for Evaluation of Translation (COMET)	text	Supervised	https://machinetranslate.org/comet#:~:text=COMET%20
33	BERTscore	text	Supervised	https://arxiv.org/abs/1904.09675
34	Statistical Parity Difference (SPD)	tabular	All	https://arxiv.org/abs/1104.3913
35	Character error rate (CER)	text	Supervised	https://readcoop.eu/glossary/character-error-rate-cer/
36	Mean Average Precision (MAP)	tabular	Supervised	https://www.v7labs.com/blog/mean-average-precision#:~:text=Mean%20Average%20Precision(mAP)%20is%20a%20metric%20used%20to%20evaluate,values%20from%200%20to%201.
37	Gender-based Illicit Proximity Estimate (GIPE)	tabular	All	https://arxiv.org/abs/2006.01938
38	Equal outcomes	tabular	All	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6594166/
39	Cross-lingual TRansfer Evaluation of Multilingual Encoders for Speech (XTREME-S)	text	Supervised	https://github.com/google-research/xtreme

40	System output Against References and against the Input sentence (SARI)	text	Supervised	https://huggingface.co/spaces/evaluate-metric/sari
41	Spearman's rank correlation coefficient (SRCC)	tabular	Supervised	https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient#:~:text=%2C%20is%20a%20nonparametric%20measure%20of,described%20using%20a%20monotonic%20function.
42	Pearson correlation coefficient (PCC)	tabular	Supervised	https://libguides.library.kent.edu/SPSS/PearsonCorr
43	MAUVE	text	Unsupervised	https://arxiv.org/abs/2102.01454
44	Matthews Correlation Coefficient	tabular	Supervised	https://towardsdatascience.com/the-best-classification-metric-youve-never-heard-of-the-matthews-correlation-coefficient-3bf50a2f3e9a
45	Mean Absolute Error (MAE)	tabular	Supervised	https://en.wikipedia.org/wiki/Mean_absolute_error
46	Google BLEU (GLEU)	text	Supervised	https://www.nltk.org/api/nltk.translate.gleu_score.html
47	FrugalScore	text	Supervised	https://arxiv.org/abs/2110.08559
48	Competition MATH	text	Supervised	https://huggingface.co/spaces/evaluate-metric/competition_math
49	chrF	text	Supervised	https://aclanthology.org/W15-3049.pdf
50	3D Pose Correct Keypoints	imaging	Supervised	https://openaccess.thecvf.com/content_iccv_2017/html/Zhou_Towards_3D_Human_ICCV_2017_paper.html
51	Absolute Relative Error (ARE)	imaging	Supervised	https://mathworld.wolfram.com/AbsoluteError.html
52	Average Dice coefficient	imaging	Supervised	https://www.jstor.org/stable/1932409
53	False Acceptance Rate (FAR)	time-series	Supervised	https://www.techopedia.com/definition/27569/false-acceptance-ratio-far
54	False Rejection Rate (FRR)	time-series	Supervised	https://www.webopedia.com/definitions/false-rejection/
55	Peak Signal-to-Noise Ratio (PSNR)	imaging	Supervised	https://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio
56	Structural Similarity Index (SSIM)	imaging	Supervised	https://en.wikipedia.org/wiki/Structural_similarity#:~:text=The%20structural%20similarity%20index%20measure,the%20similarity%20between%20two%20images
57	Fréchet Inception Distance (FID)	imaging	Unsupervised	https://papers.nips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html
58	Learned Perceptual Image Patch Similarity (LPIPS)	imaging	Supervised	https://github.com/richzhang/PerceptualSimilarity
59	Natural Image Quality Evaluator (NIQE)	imaging	Unsupervised	https://nl.mathworks.com/help/images/ref/niqe.html
60	Multi-Object Tracking Accuracy (MOTA)	imaging	Supervised	https://pub.towardsai.net/multi-object-tracking-metrics-1e602f364c0c

61	Higher order tracking accuracy (HOTA)	imaging	Supervised	https://www.researchgate.net/publication/345343240_HOTA_A_Higher_Order_Metric_for_Evaluating_Multi-object_Tracking
62	Kendall rank correlation coefficient (KRCC)	tabular	Supervised	https://www.jstor.org/stable/2332226
63	Frames Per Second (FPS)	Imaging	All	https://towardsdatascience.com/no-gpu-for-your-production-server-a20616bb04bd
64	Normalized Scanpath Saliency (NSS)	imaging	Unsupervised	https://arxiv.org/pdf/1604.03605.pdf
65	Kullback-Leibler Divergence (KLD)	tabular	Unsupervised	https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence
66	Cohen's Kappa coefficient	tabular	Supervised	https://en.wikipedia.org/wiki/Cohen%27s_kappa#:~:text=Cohen's%20kappa%20coefficient%20
67	Dice score	tabular	Supervised	https://en.wikipedia.org/wiki/S%C3%B8rensen%E2%80%93Dice_coefficient
68	Hamming distance	text	Unsupervised	https://en.wikipedia.org/wiki/Hamming_distance
69	Root Mean Squared Error (RMSE)	tabular	Supervised	https://en.wikipedia.org/wiki/Root-mean-square_deviation
70	Tree Edit Distance (TED)	text	Unsupervised	https://www.cic.ipn.mx/~sidorov/sngrams_ted_2015.pdf
71	Mean rank	tabular	Supervised	https://en.wikipedia.org/wiki/Knowledge_graph_embedding#Mean_rank_(MR)
72	Mean of Predicted Reciprocal Ranks (MRR)	tabular	Supervised	https://en.wikipedia.org/wiki/Knowledge_graph_embedding#Mean_reciprocal_rank_(MRR)
73	Normalized Discounted Cumulative Gain (NDCG)	tabular	Supervised	https://arize.com/blog-course/ndcg/
74	Normalized Mutual Information (NMI)	tabular	Unsupervised	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.normalized_mutual_info_score.html#:~:text=Normalized%20Mutual%20Information%20(NMI)%20is,and%201%20(perfect%20correlation).
75	Normalized Power Spectrum Similarity (NPSS)	time-series	Unsupervised	https://openaccess.thecvf.com/content_CVPR_2019/papers/Gopalakrishnan_A_Neural_Temporal_Model_for_Human_Motion_Prediction_CVPR_2019_paper.pdf
76	Percentage of correct keypoints (PCK)	imaging	Supervised	https://www.v7labs.com/blog/human-pose-estimation-guide
77	Scale-invariant signal-to-distortion ratio improvement (SI-SDRi)	imaging	Supervised	https://ieeexplore.ieee.org/abstract/document/8937253
78	Inferred Average Precision (infAP)	tabular	Supervised	https://www-nlpir.nist.gov/projects/tv2006/infap/inferredAP.pdf
79	nuScenes Detection Score (NDS)	tabular	Supervised	https://openaccess.thecvf.com/content_CVPR_2020/papers/Caesar_nuScenes_A_Multimoda

				l_Dataset_for_Autonomous_Driving_CVPR_2020_paper.pdf
80	Normalized Mean Error (NME)	tabular	Supervised	https://bmvc2019.org/wp-content/uploads/papers/0772-paper.pdf
81	SAFE (Sustainable, Accurate, Fair and Explainable & Interpretable)	tabular	All	https://bancaria.it/en/livello-2/archive-2/last-summary/april-2022/safe-ai-sustainable-accurate-fair-and-explainable-artificial-intelligence-in-finance/
82	Global Feature Importance Spread (GFIS)	tabular	All	https://arxiv.org/pdf/2302.12094.pdf
83	Local Feature Importance Spread Stability (LFISS)	tabular	All	https://arxiv.org/pdf/2302.12094.pdf
84	Predictions Groups Contrast (PGC)	tabular	All	https://arxiv.org/pdf/2302.12094.pdf
85	α-Feature Importance (αFI)	tabular	All	https://arxiv.org/pdf/2302.12094.pdf
86	Partial Dependence Complexity (PDC)	tabular	All	https://arxiv.org/pdf/2302.12094.pdf
87	Surrogacy Efficacy Score (SESc)	tabular	All	https://arxiv.org/pdf/2302.12094.pdf
88	Data Shapley	tabular	All	https://arxiv.org/abs/1904.02868
89	Beta Shapley	tabular	All	https://arxiv.org/abs/2110.14049
90	Data Banzhaf	tabular	All	https://arxiv.org/abs/2205.15466
91	CLIPSBERTScore	tabular	Supervised	https://arxiv.org/abs/2211.02580
92	Variable Importance Cloud (VIC)	tabular	All	https://arxiv.org/abs/1901.03209
93	Shapley Variable Importance Cloud (ShapleyVIC)	tabular	All	https://arxiv.org/abs/2110.02484
94	Local Interpretable Model-agnostic Explanation (LIME)	tabular	All	https://arxiv.org/abs/1602.04938
95	Shapley Additive Explanation (SHAP)	tabular	All	https://dl.acm.org/doi/10.5555/3295222.3295230
96	Local Explanation Method using Nonlinear Approximation (LEMNA)	tabular	All	https://dl.acm.org/doi/10.1145/3243734.3243792
97	Contextual Outlier Interpretation (COIN)	tabular	All	https://arxiv.org/abs/1711.10589
98	Rank-Aware Divergence (RADio)	tabular	All	https://arxiv.org/abs/2209.13520
99	Conditional Demographic Disparity (CDD)	tabular	All	https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-data-bias-metric-cddl.html
100	Hellinger Distance	tabular	Unsupervised	https://ieeexplore.ieee.org/document/9686689
101	SAFE Artificial Intelligence in finance	tabular	All	https://doi.org/10.1016/j.frl.2023.104088
102	SVEva Fair	tabular	All	https://arxiv.org/abs/2107.12049
103	WinoST	tabular	All	https://aclanthology.org/2022.lrec-1.230.pdf
104	Log odds-ratio	tabular	All	https://en.wikipedia.org/wiki/Odds_ratio

105	Earth Mover's Distance	tabular	Unsupervised	https://en.wikipedia.org/wiki/Earth_mover%27s_distance

3.6.4 Metrics for data privacy assessment

Privacy risk metrics aim to assess how vulnerable a dataset is to privacy breaches, after anonymization, by assessing a variety of key metrics, including [125]: (i) the **individual risk** which calculates the re-identification risk for each record by assessing the size of groups formed by quasi-identifiers, (ii) the **global risk** which is a dataset-wide measure that aggregates the individual risks and offers an average re-identification risk across the dataset, (iii) the **population unique** which identifies records that are unique within the dataset based on quasi-identifiers, (iv) the **k-anonymity** which measures the anonymity level by identifying the smallest group size formed by quasi-identifiers, (v) the **l-diversity** which evaluates the variety of sensitive attribute values within quasi-identifier groups, (v) the **t-closeness** which compares the distribution of sensitive attributes within quasi-identifier groups to the overall dataset, and (vi) the **δ-presence** which calculates the probability that an individual from the original dataset remains in the anonymized dataset.

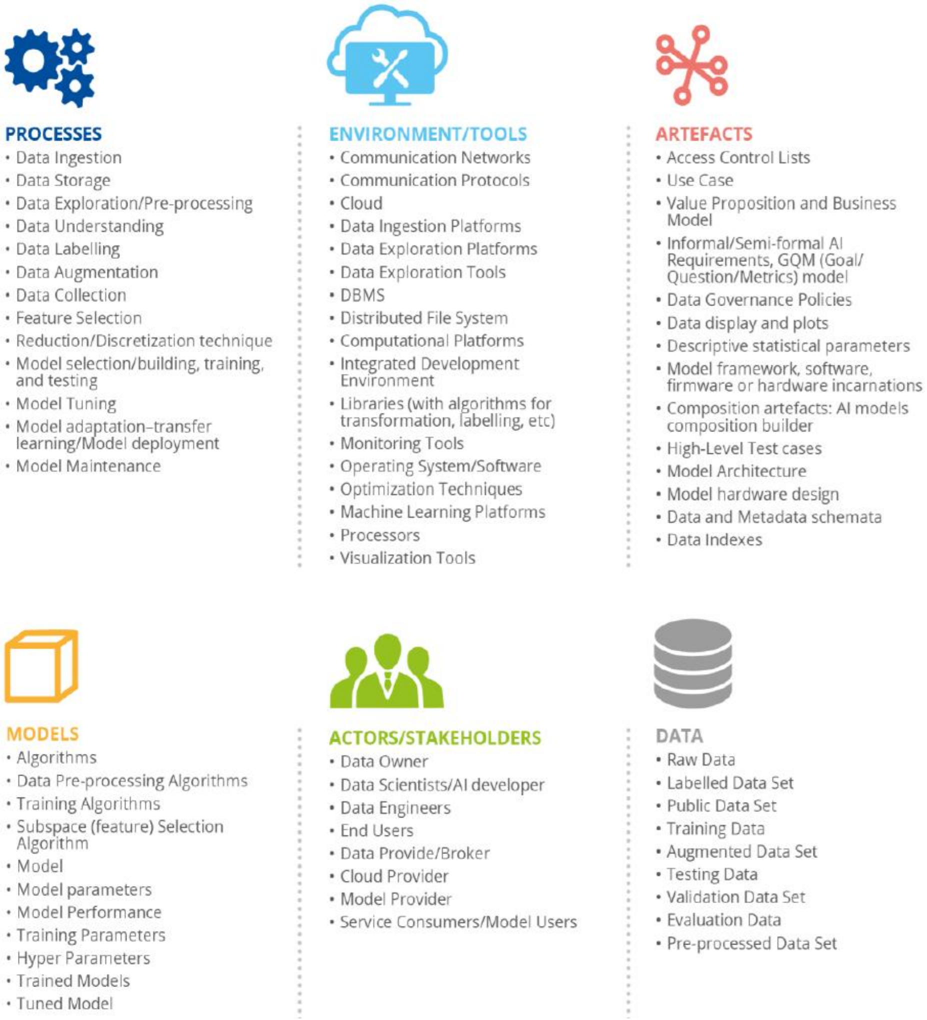
Several **data utility metrics** are also used to evaluate the loss of data in the anonymization process (due to data distortion), including [125]: (i) the **data utility loss** which measures how much the anonymized data deviates from the original using Root Mean Square Error (RMSE), with higher values indicating greater distortion, (ii) the **information loss** which evaluates the degree of information loss by calculating the mutual information score. A lower score signifies a greater loss in the relationships and structure within the data, (v) the **data comparison** which compares statistical summaries, such as mean and standard deviation, between the original and anonymized datasets, and (vi) the **disclosure risk score** which calculates the percentage of real records that exactly match the anonymized data.

4 The FAITH AI_TAF - v01

Here we present the first version of the FAITH AI_TAF to assess all assets of the AI system. All phases of the framework will be presented here. The framework can be used to assess all types of threats i.e. social, behavioural, technical, cognitive, legal, and ethical and provide measurements for all individual stages of the AI system lifecycle. Mitigation actions (technical and social countermeasures/controls) will be presented for all types of threats.

4.1 Scope of the FAITH AI_TAF, Principles and Assumptions

The FAITH AI Trustworthiness Assessment Framework (FAITH AI_TAF) is designed specifically to manage the trustworthiness risks associated with AI systems by assessing and optimizing a set of trustworthiness characteristics, such as accuracy, robustness, fairness, and transparency, among others.



PROCESSES
- Data Ingestion
- Data Storage
- Data Exploration/Pre-processing
- Data Understanding
- Data Labelling
- Data Augmentation
- Data Collection
- Feature Selection
- Reduction/Discretization technique
- Model selection/building, training, and testing
- Model Tuning
- Model adaptation–transfer learning/Model deployment
- Model Maintenance

ENVIRONMENT/TOOLS
- Communication Networks
- Communication Protocols
- Cloud
- Data Ingestion Platforms
- Data Exploration Platforms
- Data Exploration Tools
- DBMS
- Distributed File System
- Computational Platforms
- Integrated Development Environment
- Libraries (with algorithms for transformation, labelling, etc)
- Monitoring Tools
- Operating System/Software
- Optimization Techniques
- Machine Learning Platforms
- Processors
- Visualization Tools

ARTEFACTS
- Access Control Lists
- Use Case
- Value Proposition and Business Model
- Informal/Semi-formal AI Requirements, GQM (Goal/ Question/Metrics) model
- Data Governance Policies
- Data display and plots
- Descriptive statistical parameters
- Model framework, software, firmware or hardware incarnations
- Composition artefacts: AI models composition builder
- High-Level Test cases
- Model Architecture
- Model hardware design
- Data and Metadata schemata
- Data Indexes

MODELS
- Algorithms
- Data Pre-processing Algorithms
- Training Algorithms
- Subspace (feature) Selection Algorithm
- Model
- Model parameters
- Model Performance
- Training Parameters
- Hyper Parameters
- Trained Models
- Tuned Model

ACTORS/STAKEHOLDERS
- Data Owner
- Data Scientists/AI developer
- Data Engineers
- End Users
- Data Provide/Broker
- Cloud Provider
- Model Provider
- Service Consumers/Model Users

DATA
- Raw Data
- Labelled Data Set
- Public Data Set
- Training Data
- Augmented Data Set
- Testing Data
- Validation Data Set
- Evaluation Data
- Pre-processed Data Set

Figure 7: AI assets (ENISA 2020).

Main Principles and Assumptions

The FAITH AI_TAF will be used to identify and estimate threats (social and technical), vulnerabilities and consequences (likelihood of threats), estimate risks and propose mitigation measures and controls for all components of the AI system (e.g., data, trained models, AI participants, algorithmic pipelines) under assessment. FAITH AI_TAF is based on the following principles:

- ***Inclusive:*** Incorporates past experiences on risk assessment and builds upon them.
- ***Human centric:*** The trustworthiness of the AI participants is taken into account.
- ***Holistic:*** Applicable to each phase of the AI lifecycle.
- ***Tool independent****:* Independent of tools used.
- ***Agile:*** Can be used by all sectors (by adjusting the responses).
- ***Risk Assessment Standard based****:* Compliant with ISO27005 [23], ISO42001 (AI risk management) [145].
- ***Versatility:*** Beneficial to AI stakeholders regardless of the industry/sector.
- ***Global Reach*****:** Compliant with European and international initiatives, standards, and guidelines (e.g. ENISA FAICP, CEN/CENELEC, ETSI SAI, NIST AI RMF).

FAITH AI_TAF adopts and extends the ENISA Framework for AI Cybersecurity Practices (FAICP) [9] approach which views the AI systems as part of an ICT operational infrastructure:

Layer I (foundations/assumptions): FAITH AI_TAF makes the following assumptions

- Cybersecurity practices are implemented across all ICT environments (certified by ISO27001, or they follow NIST or ENISA best practices for security management) involved in hosting, operating, developing, integrating, maintaining, supplying the AI system (under assessment).
- The focus is on threats affecting the Trustworthy Characteristics. Cybersecurity of the ICT environment is considered when specifically associated with elements related to AI and it is assumed generally addressed by known best practices.
- Data sources are assumed to be secure (they are certified/tested) (and not generate poisoned data).
- Each AI asset of the AI system under assessment (SUT) has an owner (who is responsible for the asset).
- The responsibility for conducting a trustworthiness risk assessment process using FAITH AI_TAF lies with the Risk Assessor or the ISMS coordinator of the organization (it is the organization’s decision to select one person). All AI participants may be involved in the process, however in this 1st version of the FAITH AI_TAF only one person can provide the responses.
- We assume that no cascaded/propagated threats are feasible i.e., the AI system is either isolated or any interdependent components are secure (i.e., implemented controls do not enable the propagation of their vulnerabilities).

Layer II (FAITH AI_TAF)-General AI trustworthiness assessment: The unique characteristics of AI components throughout their lifecycle, including their properties, potential threats, AI participants and necessary mitigation actions and controls for all dimensions of trustworthiness are considered here. The consequences (impacts)of the threats to the dimensions of trustworthiness are evaluated. The risks will be evaluated for the general AI threats for the AI components.

The trustworthiness assessment is human-centric. This entails that the trustworthiness characteristics for which assessments will be conducted will be selected to fit the needs and requirements of users and stakeholders, as well as the context of use. Assessment of trustworthiness characteristics, such as those concerning the reliability and validity of the system under test, may require careful specification and refinement to adequately fit the needs of the users and stakeholders. Furthermore, user and stakeholder perspectives may be required in situations where optimization of trustworthiness requires compromises between trustworthiness characteristics.

The risk management process occurs at every phase of the AI lifecycles.

Layer III (FAITH AI_TAF in LSPs)- Environmental trustworthiness assessment: The following business factors will further customise the assessment:

- Sector that the AI system operates in
- Criticality and Intended use(s) of the AI system
- AI teams involved in AI life cycle
- Trustworthiness dimensions relevant to the AI system
- Risk appetite (the amount and type of trustworthiness risk the organization is willing to accept in pursuit of its business objectives. It defines the organization’s tolerance for uncertainty and potential losses while balancing opportunities for growth and innovation).

These factors may modify the risk evaluations (in Layer 2). The FAITH LSPs will be analysed to reveal the realistic risk estimates for each environment (intended use(s), participant, business objectives).

The FAITH AI_TAF will be used to assess the threats and estimate the risks of all components of the AI system during its entire life cycle against these threats considering sectoral characteristics and requirements, assuming that the infrastructure that hosts the AI system is secure (e.g. ISO27001 certified) , data sources certified (e.g. ISO15408 certified) and any risk of interconnected systems are treated .

The application of the FAITH AI_TAF on the range of LSPs included in the FAITH project implies a need to ensure that the assessment is adapted to the specific sector. At the same time the application of the FAITH AI_TAF across sectors ensures that the developed framework is sufficiently flexible so as to fit the requirements of diverse sectors. Furthermore, the range of sectors represented in the FAITH LSPs allows for a robust validation of the framework.

4.2 Phases of the FAITH AI_TAF

Given the above mentioned, the main objective is to classify a broad range of threats that encompass all aspects of trustworthiness, utilizing extensive classification efforts that incorporate insights from influential organizations such as ENISA, NIST, OWASP, MITRE, and others. Furthermore, we seek to identify the current obstacles in fully understanding the complete AI threat landscape across all dimensions of trustworthiness and emphasize the research efforts needed to tackle these challenges effectively.

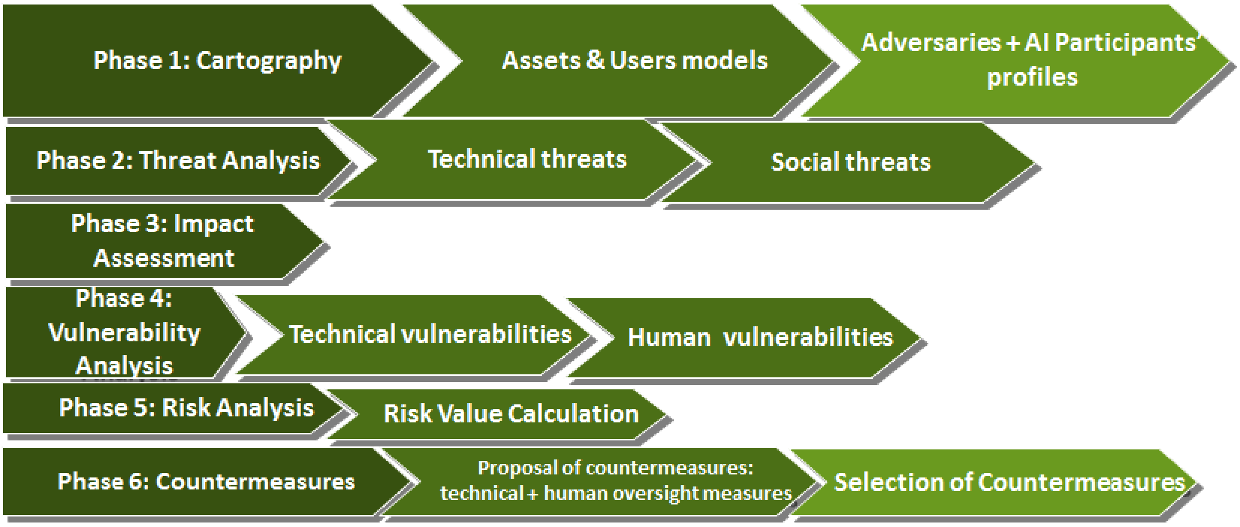

Phase 1: Cartography
Assets & Users models
Adversaries + AI Participants' profiles
Phase 2: Threat Analysis
Technical threats
Social threats
Phase 3: Impact Assessment
Phase 4: Vulnerability Analysis
Technical vulnerabilities
Human vulnerabilities
Phase 5: Risk Analysis
Risk Value Calculation
Phase 6: Countermeasures
Proposal of countermeasures: technical + human oversight measures
Selection of Countermeasures

Figure 8: FAITH AI_TAF Phases.

4.2.1 Phase 1: Cartography (setting boundaries)-Initialization Phase:

This is a critical initial stage in the development of a robust framework. During this phase, the boundary of the assessment is defined. The AI system under assessment needs to be identified, the AI system use/purpose, the AI team (participants in the AI lifecycle) and its components (AI assets).

The taxonomy as described by ENISA (Figure 8) can be used to easily classify the assets of the AI system under assessment.

An *asset model* (see example in Figure 9) of the classified AI assets will then be developed which is a structured framework used to reveal interconnections and interrelations aiming the AI assets (e.g., training models, algorithms, data, processes, AI participants) across different stages of the AI lifecycle. This model helps to ensure a comprehensive understanding of their interrelations and identifies which components require assessment. By defining the relevant assets, the asset model facilitates the prioritization of trustworthiness concerns.

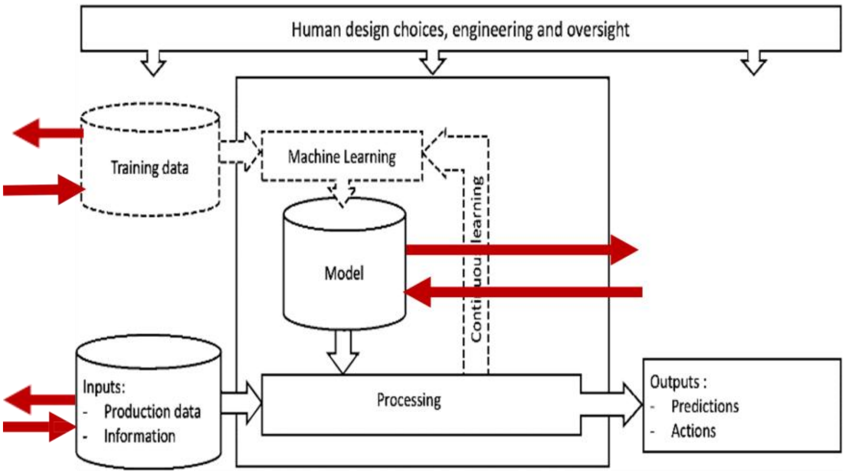


Figure 9: AI asset model of the AI system under assessment [145].

In this first version of FAITH AI_TAF, assess the risks of the individual AI assets for all its threats in isolation (no interconnected threats are considered). An asset list is kept and divided according to the different stages of the AI lifecycle. By defining the relevant assets, the relevant threats, vulnerability and controls can be identified.

Every AI asset has an owner (the AI team responsible for the asset). Please note that an asset can have 1 team of participants as owners (in entities that host large scale infrastructure).

The criticality of the AI system will depend upon:

⇒ if the AI system falls under the qualification of a high risk AI-system as established by Article 6 of the AI Act, such as the high risks systems listed in **AI Act** Annex III

⇒ If the AI system poses a significant risk of harm to the health, safety or fundamental rights of natural persons, including by not materially influencing the outcome of decision making.

⇒ if the AI system is used by an Operator of Essential services (OES) and the AI system is used for the provision of an essential service as defined in Annex I of NIS 2 Directive or an important service as defined in Annex 2 of NIS2 .

⇒ If the AI system is processing personal data (where GDPR needs to apply).

Table 18: Criticality level of an AI system.

Criticality level	Conditions for defining criticality level
Very High	if the AI system falls under the qualification of a high risk AI-system as established by Article 6 of the AI Act, such as the high risks systems listed in **AI Act** Annex III If the AI system poses a very high risk of harm to the health, safety or fundamental rights of natural persons, including by not materially influencing the outcome of decision making.
High	if the AI system poses a high risk of harm to the health, safety or fundamental rights of natural persons, including by not materially influencing the outcome of decision making

	and /or If the Ai system is used by an Operator of Essential services (OES) and the AI system is used for the provision of an **essential service** as defined in Annex I of NIS 2 Directive
Substantial	If the AI system poses substantial risk of harm to the health, safety or fundamental rights of natural persons, including influencing the outcome of decision making If the AI system is used by an Operator of Essential services (OES) and the AI system is used for the provision of an **important service** as defined in Annex 2 of NIS2
Medium	If the AI system poses medium risk of harm to the health, safety or fundamental rights of natural persons, including influencing the outcome of decision making If the AI system is for the provision of an **important service** as defined in Annex 2 of NIS2
Low	If the AI system poses low or no risk of harm to the health, safety or fundamental rights of natural persons, including influencing the outcome of decision making

All scales proposed in the FAITH AI_TAF can be adjusted according to the criticality level of the AI system.

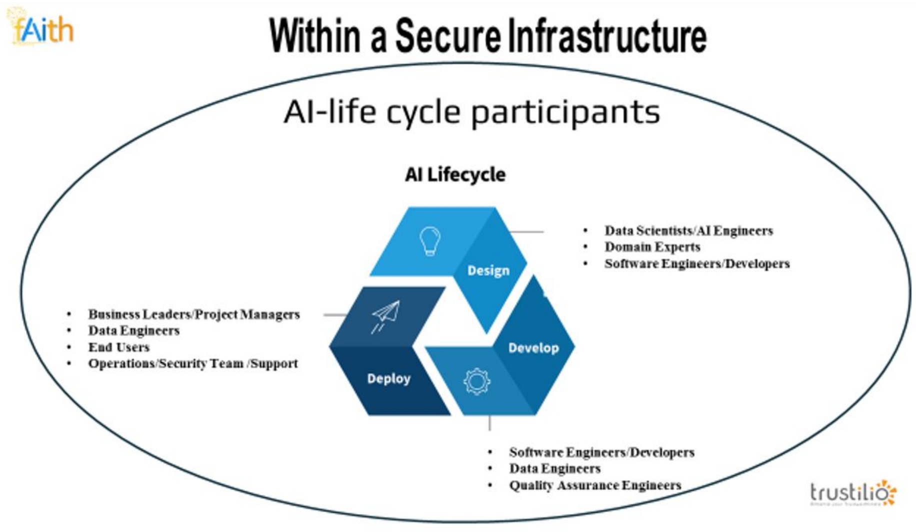

fAith
Within a Secure Infrastructure
AI-life cycle participants
AI Lifecycle
Design
Develop
Deploy
Data Scientists/AI Engineers
Domain Experts
Software Engineers/Developers
Business Leaders/Project Managers
Data Engineers
End Users
Operations/Security Team /Support
Software Engineers/Developers
Data Engineers
Quality Assurance Engineers
trustilio

Figure 10: AI user (participants) model.

Concurrently, a *user (organization team with AI participants) model* will be created, detailing the various AI teams involved in all phases of the AI system under assessment, their roles, and access levels, which helps in mapping out legitimate interactions within the system.

Organizational aspects – Maturity for Teams for handling AI trustworthiness

The **Trustworthy AI Maturity for Teams** for adopting effective practices in identifying AI trustworthiness threats, mitigating AI trustworthiness risks will be measured here, including the understanding of the roles and behaviours of AI-enabled teams and entities interacting within the environment. **In this first version (v1) of the FAITH AI_TAF, we assume that only the AI asset owner's team is utilizing the asset, and therefore, we estimate the trustworthiness of each organizational team rather than individual participants**. To capture the varying levels of expertise among AI users, **two separate but closely related questionnaires are proposed**. The only distinction between them, lies in the Technical Proficiency Questions which is adjusted to the responder's roles: Technical Users and AI Domain Users as these can be seen below in the relevant section.

There is an option also to evaluate potential **adversaries' sophistication (tA)** if the organization has such cybersecurity intelligence gathered (from past incidents, collaboration with ISACS, CERTS)**.** This optional assessment involves analysing the capabilities, goals, and characteristics of adversaries who might attempt to exploit the system, based on past experience and historical data.

The benefit of assessing the trustworthiness of the AI participants may be particularly relevant for some sectors or application domains, and less relevant for others. For example, in sectors or contexts with a broad range of AI users that have not undergone initial filtering or assessment, such assessments may be particularly useful. In sectors or contexts with trained and selected personnel, requirement assessments may already be in place through organizational measures and there is, hence, not required to conduct this specifically for an AI trustworthiness assessment. In such cases, an assessment of AI readiness at a team (e.g. risk assessment team) or organizational level may be more relevant.

Measurement of **Trustworthy AI Maturity for Teams**

The maturity of the AI team will be evaluated by the AI risk manager by responding to the following statements as (Strongly Disagree = 1, Disagree = 2, Neutral = 3, Agree = 4, Strongly Agree = 5:

Proactivity and Threat Awareness

- The team understands the technological, social and compliance requirements of the multidimensional aspects of AI trustworthiness (cybersecurity, privacy, quality, robustness, transparency, explicability etc).
- The team routinely identifies at potential technological, operational or social AI threats.
- In scenarios where an AI threat is exploited or an AI incident occurs, the team acts to mitigate it in line with the requirements of their quality management system.

- The team understands the organization's security and AI policy, including its commitments and objectives; the team is aware of the quality objectives that apply to their specific roles and responsibilities; and understands how nonconformities can negatively affect the AI operations and how these impact their business.

Responsibility and Ethics

- The team collectively ensures that all members understand their roles in maintaining AI trustworthiness, holding routine review sessions as part of internal quality reviews.
- The team consistently prioritizes adherence to AI trustworthiness best practices, directives, standards and guidelines, even during high-pressure scenarios.
- The team knows the intended use of the AI systems that they operate, their normal operation and their expected outcomes.

Innovation and Adaptability

- The team has an established routine for implementing or enhancing new mitigation actions (e.g., technological control, policy, procedure) to address AI trustworthiness challenges creatively.
- If faced with a significant error, the team collectively develops a revised process to prevent recurrence, shared in accordance with the requirements of the quality management system.

Resilience

- The team ensures effective recovery and organizational business continuity within the first 24 hours after an incident, consistently meeting project deadlines and maintaining a success rate above 90%.
- Technological failures do not lead to a drop in team performance metrics. (Reverse-coded).

Collaboration

- The team collaborates effectively and has an established routine for sharing new key insights or data points that may address technological threats.
- The team builds professional relationships with internal and external partners, encouraging meetings to enhance coordination and enhance the threat intelligence.

Integrity

- The team consistently upholds ethical principles, legal compliance and adheres to professional codes of conduct in its operations.

Technical Proficiency (Questionnaire 1: Technical Users)

- The team demonstrates proficiency in managing data quality (e.g. data wrangling, distributed databases for handling large datasets, understanding how to create high-quality, unbiased synthetic datasets) by conducting routine audits of datasets and their use.
- The team applies advanced technological tools (e.g. optimising AI models; knowledge and tools to ensure model transparency, interpretable models, protected models from adversarial attacks; reduction of algorithmic biases, privacy, auditability, robustness) in ongoing projects where this is required.

Technical Proficiency (Questionnaire 2: AI domain Users)

- The team can critically assess AI-generated results, identifying inconsistencies, biases, or errors that may consequence decision-making.
- The team understands and applies basic AI reliability and safety practices, such as verifying data sources, interpreting AI outputs, and following ethical guidelines.

Problem Solving

- The team is skilled in resolving issues, completing 90% of identified challenges through interdisciplinary collaboration.

Resource Accessibility

- The team has access to high-performance computing tools and networks, routinely engaging in sessions with external experts to enhance capabilities.
- Limited interaction with external technological communities is detrimental to team progress. (Reverse-coded).

Policy Adherence

- The team adheres to policies by maintaining a compliance score on Trustworthy AI above 95% during regular audits.

Motivation and Commitment

- The team consistently demonstrates a commitment to trustworthy AI by organizing regular ethical reviews and discussions and attend professional trainings.

Privacy and Compliance

- The team prioritizes privacy and legal compliance for trustworthy AI by achieving at least 95% adherence in internal audits.

Openness to Interventions

- The team welcomes external feedback, routinely attending training sessions annually to refine practices.

- Resistance to changes in workflows that enhance trustworthiness is a challenge. (Reverse-coded).

Scoring and Interpretation

Each question is scored on a Likert scale (Strongly Disagree = 1, Disagree = 2, Neutral = 3, Agree = 4, Strongly Agree = 5).

The calculation methodology ensures that individual responses contribute to a collective organizational score. The process involves the following steps:

1. Weight Assignment: Each dimension of trustworthiness is assigned a weight based on its criticality to the organization. Example weights:

- Responsibility and Ethics: 25%
- Technical Proficiency: 20%
- Collaboration: 20%
- Proactivity and Threat Awareness: 15%
- Privacy and Compliance: 20%

2. Response Scoring: Each question is rated on a 5-point Likert scale. Reverse-coded items are adjusted to ensure consistency (e.g., a "1" for a reverse-coded question is converted to a "5").

3. Dimension Scores: For each dimension (e.g., Ethics, Collaboration), individual responses are averaged to create a dimension-level score.

4. Weighted Scores: The dimension-level scores are multiplied by their respective weights. For example:

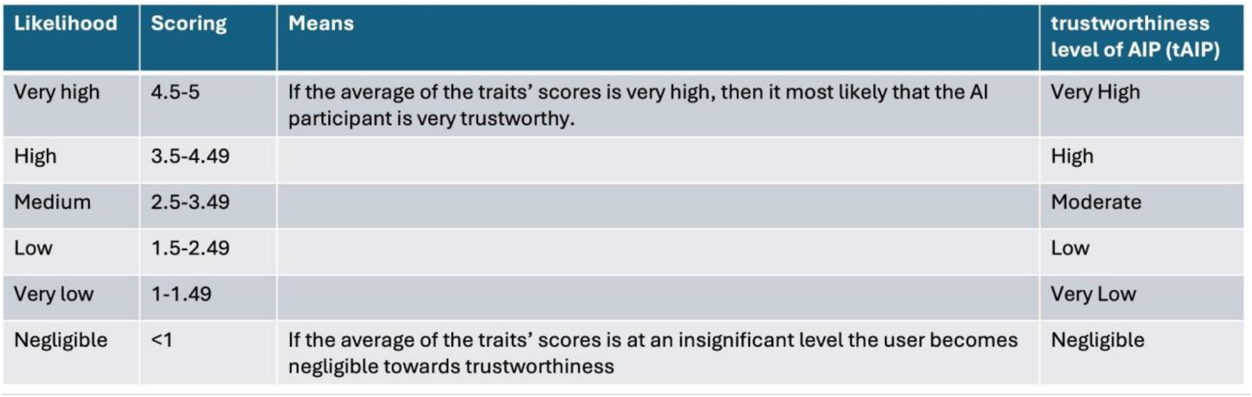
- Ethics Score = (Average Ethics Responses) × 0.25
- Collaboration Score = (Average Collaboration Responses) × 0.20

5. Normalization: To standardize the scores, a logarithmic transformation is applied if the data distribution shows significant skewness. This step ensures comparability across dimensions.

6. Overall Organizational Score: The weighted scores for all Questions are summed to produce the overall organizational trustworthiness score.

7. Categorization: The final score is categorized into trustworthiness levels using predefined thresholds:

4.5 - 5: Very High, 3.5 - 4.49: High, 2.5 - 3.49: Moderate, 1.5 - 2.49: Low, 1 - 1.49: Very Low, > 1: Negligible. The scores are used for the trustworthiness estimation (Figure 11), where Likelihood is the likelihood of an AIP score highly in most of the traits. We use direct referral to the relationship between likelihood and trustworthiness e.g the likelihood of the presence of many traits is directly linked to level of trustworthiness and the tP is inverse to the risk level.



Likelihood	Scoring	Means	trustworthiness level of AIP (tAIP)
Very high	4.5-5	If the average of the traits' scores is very high, then it most likely that the AI participant is very trustworthy.	Very High
High	3.5-4.49		High
Medium	2.5-3.49		Moderate
Low	1.5-2.49		Low
Very low	1-1.49		Very Low
Negligible	<1	If the average of the traits' scores is at an insignificant level the user becomes negligible towards trustworthiness	Negligible

Figure 11: Trustworthiness of AIP estimate

The overall score can then be utilized by the organization by following the mitigation recommendations of the TrustSense scale (Table 19):

Table 19: Mitigation recommendations of the TrustSense scale.

Likelihood	**Scoring**	**Interpretation of Results**	**AI Maturity for Teams**	**Mitigation Recommendations**
Very High	4.5 - 5	The team demonstrates consistently high maturity regarding trustworthy AI	Very High	To maintain this level, organize regular team training sessions, recognize collective achievements, and promote a culture of continuous improvement.
High	3.5 - 4.49	The team largely adheres to requirements for trustworthy AI, with minor areas for improvement.	High	Enhance organizational training programs, encourage cross-team collaborations, and refine adherence to ethical codes and organizational policies to elevate performance.
Medium	2.5 - 3.49	The team shows partial adherence to requirements for trustworthy AI, indicating areas needing attention.	Moderate	Implement structured training initiatives, strengthen collaborative practices, and promote organizational mentorship to address identified gaps.
Low	1.5 - 2.49	Significant gaps in trustworthy AI maturity exists at the team's level.	Low	Facilitate intensive team workshops, prioritize ethical compliance, and establish policies to strengthen trustworthiness practices across teams.

Very Low	1 - 1.49	The team faces considerable challenges in trustworthy AI maturity.	Very Low	Commit to comprehensive retraining programs, monitor collective progress through evaluations, and establish supervised practices to rebuild foundational trustworthiness traits.
Negligible	<1		Negligible	

Measuring Sophistication of potential AI adversaries[6] (tA)

Estimation of adversaries maturity will be optional in the FAITH AI_TAF framework. Such estimations can be considered if the organisations have historic information regarding adversaries from past cybercriminal investigations (e.g. tracking digital footprints, analysing behavioural patterns, understanding motivation).

Similarly to the AI teams ' trustworthiness maturity estimation based on their profiles, we have developed and proposed a scale for profiling potential AI adversaries (in case the organisations wish to identify potential internal adversaries and their levels to conduct sophisticated attacks). Each question is rated on a 5-point Likert style scale:

1. In social gatherings, I am usually the one who initiates conversations and interactions.
2. I enjoy taking charge and leading group projects or activities.
3. I prefer a fast-paced, dynamic lifestyle with many activities.
4. I generally maintain a positive and optimistic outlook on life.
5. I am comfortable expressing my thoughts and opinions openly, even if they are controversial.
6. I am meticulous and organized in my work and personal life.
7. I am persistent in pursuing my goals, even when faced with obstacles.
8. I am disciplined and able to resist temptations that might hinder my progress.
9. I have a strong sense of responsibility and duty towards my commitments.
10. I am confident in my ability to achieve my goals and overcome challenges.
11. I have a vivid imagination and enjoy creative pursuits.
12. I am curious about scientific and intellectual topics and enjoy learning new things.
13. I am open to trying new experiences and exploring different cultures and ideas.
14. I enjoy engaging in abstract thinking and considering philosophical questions.
15. I am comfortable expressing my emotions and feelings openly.
16. I find it difficult to conform to traditional social norms and expectations.

[6] this is optional and only if there is knowledge/access of the adversaries

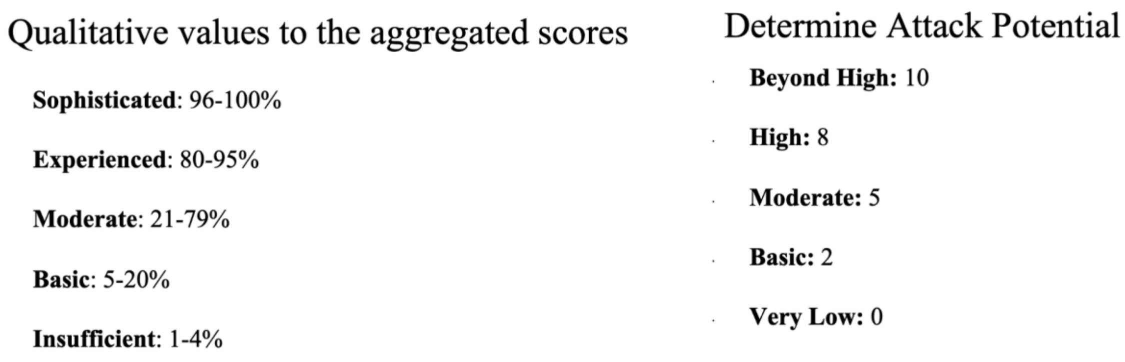
17. I am comfortable building online relationships with people I have never met in person.
18. I am more likely to form strong bonds with people in online communities than in real life.
19. I prefer to communicate online rather than in person.
20. I am skilled at manipulating people's emotions and actions online.
21. I have a strong understanding of network architectures and protocols.
22. I am proficient in various operating systems, programming languages, and software tools.
23. I am skilled at analysing and solving complex technical problems.
24. I am observant of security practices and can identify vulnerabilities in systems and behaviours.
25. I have experience using security scripts and forensic tools.
26. I have access to significant computing resources and time to dedicate to technical pursuits.
27. I have insider knowledge or access to sensitive information within an organization or system.
28. I am motivated by the pursuit of political power or influence.
29. I am motivated by personal gain, such as financial rewards or a sense of accomplishment.
30. I am motivated by a desire to expose wrongdoing or corruption.
31. I am motivated by humanitarian or activist goals, such as social justice or environmental protection.
32. I am more likely to target systems or networks with known vulnerabilities.
33. I am interested in exploiting new and untested technologies.
34. I am more likely to target organizations with weak security practices or infrastructure.
35. I am opportunistic and will take advantage of unintentional human errors.

Scoring and Interpretation

Each question is scored on a Likert scale (Strongly Disagree = 1, Disagree = 2, Neutral = 3, Agree = 4, Strongly Agree = 5). The average scores will then be categorized as follows:

4.5 - 5: Very High, 3.5 - 4.49: High, 2.5 - 3.49: Moderate, 1.5 - 2.49: Low, 1 - 1.49: Very Low, > 1 Negligible. The scores for each section are then summed and converted to a percentage. The total percentage for each section is then compared to the ranges in Table 17 to determine the attacker profile.

For example, if a respondent scores 20 out of 25 in the Technical Traits section, their percentage score would be 80%. This would fall into the "Experienced" category in Table 6, with a corresponding score of 8 (Figure 12).



Qualitative values to the aggregated scores

Sophisticated: 96-100%

Experienced: 80-95%

Moderate: 21-79%

Basic: 5-20%

Insufficient: 1-4%

Determine Attack Potential

- **Beyond High:** 10
- **High:** 8
- **Moderate:** 5
- **Basic:** 2
- **Very Low:** 0

Figure 11: Likelihood of attack potential in relation to adversaries profiles.

4.2.1.1 Output of Phase 1:

By thoroughly mapping out these components, the cartography (initialization) phase establishes a foundational blueprint of the trustworthiness assessment. The components of the AI systems under assessment, the AI participants for all stages of the AI lifecycle, the criticality level of the AI system. Asset and user (AI participants) models will be provided. AI organizational teams and potential adversaries' maturity will be calculated. The relevant trustworthiness dimensions for the system under assessment.

4.2.2 Phase 2: Threat Assessment

Threat assessment is a comprehensive examination of the diverse threats of all AI assets during each phase of the AI lifecycle that could undermine system trustworthiness. It encompasses technical threats like malware, data poisoning as well as social threats such as phishing, social engineering (depending of the trustworthy AI maturity level of the organization's teams). Identify threats related to data quality, AI model performance, and operational deployment. Appendix A provides a first version of a list of threats, controls for all assets in AI systems that can be used to identify the threats in each AI component of the AI system under assessment.

Measuring Threat levels: Frequency of Threats [7]

In this phase we also estimate the occurrence of the threat (Table 20). Factors influencing the occurrence of a threat include:

- historical data: previous occurrences of similar threats can provide insight into future probabilities;
- environmental factors: Conditions in a specific area or sector that the AI system is being used (e.g., natural disasters, political/economic/ business stability);
- stability & trends: geopolitical situations (e.g. economic crisis, disasters, wars, pandemia), technological trends (e.g. AI attack systems) can indicate rising threats.

Table 20: Occurrence of the threat.

[7] Depending upon the criticality of the sector/application/use of the AI system, this proposed will be adjusted

Threat Level	Frequency (as reported by the administrators or by logs)
Very high	Twice a year
High	Once a year
Medium	Once every 2 years
Low	Once every 5 years
Very low	Once every 10 years
Negligible	Never

The proposed threat level can be adjusted according to the criticality of the AI system and the organisations'" risk appetite" (the amount and type of risk the organization is willing to accept in pursuit of its objectives. It defines the organization's tolerance for uncertainty and potential losses while balancing opportunities for growth and innovation).

For example:

For AI systems with **criticality level Very High** the Threat level is Very high if it occurs twice in the last 5 years etc.

For AI systems with **criticality level High** the Threat level is Very high if it occurs twice in the last 3 years etc.

For AI systems with **criticality level Substantial** the Threat level is very high if it occurs twice in the last 2 years etc.

For AI systems with **criticality level Medium** the Threat level is very high if it occurs twice in the last year etc.

For AI systems with **criticality level Low** the Threat level is very high if it occurs twice a year etc.

4.2.2.1 Output of Phase 2:

All threats (technical and social) have been identified. The frequency of occurrence of each threat to the AI components of the AI system under assessment is estimated.

4.2.3 Phase 3: Consequence (Impact) Assessment

Evaluate the potential consequence of each threat to the various dimensions of trustworthiness (technical characteristics such as accuracy and robustness, transparency, cybersecurity, fairness, explainability, accountability, privacy).

Carefully evaluating potential threats and their potential effects to the various dimensions of trustworthiness in each phase of the AI lifecycle will be conducted here. It examines the consequences of threats to the various dimensions of trustworthiness. Using the OWASP, ENISA repositories we can identify the overlapping consequences of threats. For example, the threats data loss, model poisoning damage various dimensions of trustworthiness i.e. accuracy, fairness, cyber security.

Table 21: Consequence of each threat to the dimensions of trustworthiness.

Consequence (Impact) Level	Means
Very high	The threat has **very serious** consequences in the dimensions of trustworthiness. The threat consequences the dimensions in various ways
High	The threat has **serious consequences** in the dimensions of trustworthiness
Medium	The threat has **some consequences** in the dimensions of trustworthiness
Low	The threat has **low consequences** in the dimensions of trustworthiness
Very low	The threat has **very low consequences** in the dimensions of trustworthiness
Negligible	no consequences

The above Table will be adjusted according to the relevant dimensions of trustworthiness relevant to the system under assessment. The consequences to the various dimensions may vary and then we take the average.

The impact level of each threat will be further refined based on the potential consequences—technological, legal, financial, and others—that organizations may face when these trustworthiness dimensions are compromised.

4.2.3.1 Output of Phase 3:

Consequence assessments for each threat against the trustworthiness dimensions. The consequence assessments will include assessment scores. If required, the scores may be complemented with qualitative reports to detail any relevant assessment findings - e.g. to detail the nature of the treats.

4.2.4 Phase 4: Vulnerability Assessment

This phase focuses on identifying weaknesses in the AI system, including technical vulnerabilities (e.g. Software flaws, lack of data governance practices, network weaknesses) that could be exploited by attackers. It also addresses human vulnerabilities (low trustworthy AI maturity in the team).

In this phase the controls that have been implemented for each threat will be reported as % of controls over total numbers of available controls.

Available controls AI controls can be found in various knowledge data basis (DB) e.g. OWASP AI , ENISA , NIST AI Risk Management Framework (AI RMF 1.0) .

These DBs provide catalogues of AI threats, vulnerabilities and controls that assessors can use as a benchmark. Technical vulnerabilities can be identified with the utilization of AI assessment tools (see Section 3.5), penetration testing, vulnerability scans, and social engineering assessments. The FAITH AI_TAF in this phase enables organizations to identify their missing controls for each threat and measure their vulnerabilities for all threats in the AI system under use.

4.2.4.1 Measuring Vulnerability Level for each AI threat to the AI component[8]

Table 22: Vulnerability level.

Vulnerability Level	Means
Very high	None (0%) of controls have been implemented over total number of available controls for preventing the exploit of the threat
High	Very few (<20-40 %) of controls have been implemented over total number of available controls for preventing the exploit of the threat
Medium	Few (< 40-60%) of controls have been implemented over total number of available controls for preventing the exploit of the threat
Low	Many (>60-80%) of controls have been implemented over total number of available controls for preventing the exploit of the threat
Very low	Most (>80-99%) of controls have been implemented over total number of available controls for preventing the exploit of the threat
Negligible	All (100%) of controls have been implemented over total number of available controls for preventing the exploit of the threat

The proposed vulnerability level can be adjusted according to the criticality of the AI system and the organizations' "risk appetite".

For example:

For AI systems with **criticality level Very High** the vulnerability level can be **Very High** if most (< 80 - 90%) have been implemented etc.

For AI systems with **criticality level High** the vulnerability level can be **Very High** if many (< 60 - 80%) have been implemented etc.

[8] The scales can be adjusted according to the criticality of the AI system in different sectors/intended use

For AI systems with **criticality level Medium** the vulnerability level can be **Very High** if many (< 40 - 60%) have been implemented etc.

For AI systems with **criticality level Low** the vulnerability level can be **Very High** if many (< 20 - 40%) have been implemented etc.

4.2.4.2 Output of Phase 4:

Vulnerabilities of all threats to each AI component in every stage of the life cycle. Vulnerability levels have been estimated.

4.2.5 Phase 5: Risk Assessment

The risks of all AI assets of the AI system will be estimated in terms of:

- *Threat Level*: The frequency of occurrence of a threat (Phase 2).
- *Vulnerability Level*: Determine the system's susceptibility based on control measures (Phase 4).
- *Consequence Level*: Consider the potential consequences of the threat to all dimensions of trustworthiness (Phase 3).
- *Trustworthy AI maturity of teams, trustworthiness of AI participants* and *sophistication of potential adversaries* (Phase 1).

4.2.5.1 Calculating Risk levels (1st option)

A matrix-based risk calculator is a practical tool for assessing risk in AI systems by quantifying the interaction between threat level, vulnerability level, and consequence level. Here's a proposed risk calculator matrix for AI systems:

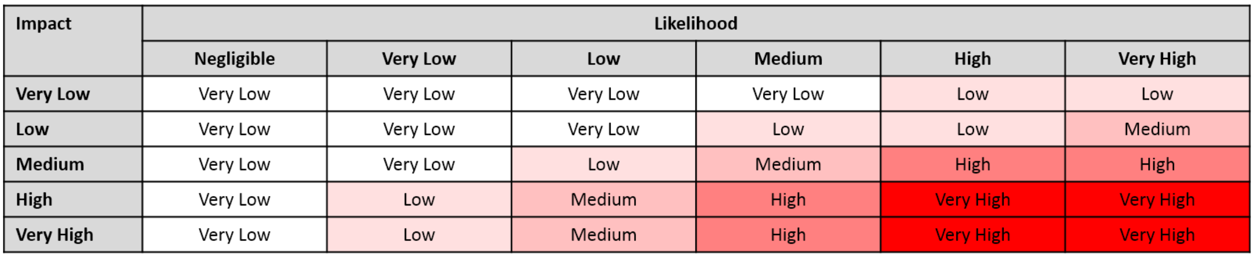
Table 23: Risk calculator matrix.

Threat Level	Vulnerability Level	Consequence Level	Risk Level
Very High	Very High	Very High	Critical
Very High	High	High	Severe
Very High	Medium	Medium	High
Very High	Low	Low	Medium
High	Very High	High	Severe
High	High	Medium	High
High	Medium	Low	Medium
High	Low	Low	Low
Medium	Very High	Medium	High
Medium	High	Medium	Medium
Medium	Medium	Low	Low
Medium	Low	Low	Minimal
Low	Very High	Low	Medium
Low	High	Low	Low
Low	Medium	Minimal	Minimal

Low	Low	Minimal	Minimal

Calculating Risk levels (2nd option)

The likelihood of a threat is the product of threat level and vulnerability level and in this case the risk matrix is the following (Figure 13):



Impact	**Likelihood**					
	Negligible	**Very Low**	**Low**	**Medium**	**High**	**Very High**
Very Low	Very Low	Very Low	Very Low	Very Low	Low	Low
Low	Very Low	Very Low	Very Low	Low	Low	Medium
Medium	Very Low	Very Low	Low	Medium	High	High
High	Very Low	Low	Medium	High	Very High	Very High
Very High	Very Low	Low	Medium	High	Very High	Very High

Figure 12: Risk Matrix.

This is base, on the ISO 27005 says risk level is a function of threat likelihood and consequence.

This version uses 5 consequence levels and 6 likelihood/TW levels, mapping to 5 risk levels the lowest likelihood always leads to the lowest available risk level. The rest of the table is the same as the original used in the risk calculator when ISO 27005 support was first introduced.

4.2.5.2 Output of Phase 5:

The risk level matrix for all threats against all AI components.

By matching these levels across the matrix, you arrive at a Risk Level that can guide action steps, such as implementing monitoring system performance, data governance systems, trainings and awareness campaigns to the AI participants.
By understanding these risks, organizations can prioritize actions to manage and reduce them effectively. This phase provides a clear basis for decision-making, helping organizations invest in measures that strengthen their overall trustworthiness and minimize potential harm.

4.2.6 Final calculation of risk levels:

The final score (independently of the option mentioned above) will consider the trustworthiness of the organizational AI maturity level (tAIP) estimated in Phase 1 and (optionally) the sophistication level (tA) of the potential adversary.

The final estimation of Risk (fR) comparing to the original Risk R, calculated above for each AI asset against each threat will be calculated as follows:

1) If we have estimated the organizational team AI maturity (tAIP) of the specific AI asset the calculation could be as follows [AI Participants in this case are defined as the AI asset Owner and users]:

fR= R- 1 if tAIP>= medium

fR= R+1 if tAIP<medium

[where 1= one level of the scale, for example from very high to high or from low to medium]

As shown partially in Table 22:

Table 24: Final Risk estimation-1.

Risk Level (R)	tAIP	Final Risk Level (fR)
Very High	Very High	High
Very High	High	High
Very High	Medium	High
Very High	Low	Very High
High	Very High	Medium
High	High	Medium
High	Medium	Medium
High	Low	Very High
Medium	Very High	Low
Medium	High	Low
Medium	Medium	Low
Medium	Low	High
Low	Very High	Very Low
Low	High	Very Low
Low	Medium	Very Low
Low	Low	Medium

2) If we also know the sophistication level of the potential adversary (tA) then:

fR=R if tA=tAIP

fR= R - 1 if tAIP>tA

fR= R + 1 if tAIP<tA

[where 1= one level of the scale, for example from very high to high or from low to medium]

As shown in Table 23 as an example:

Table 25: Final Risk estimation-2.

Risk Level (R)	tAIP	tA	Final Risk Level (fR)
Very High	Very High	Very High	Very High

Very High	High	High	Very High
Very High	Medium	Medium	Very High
Very High	Low	Low	Very High
High	Very High	High	Medium
High	High	Medium	Medium
High	Medium	Low	Medium
High	Low	Low	High
Medium	Very High	Medium	Low
Medium	High	Medium	Low
Medium	Medium	Low	Low
Medium	Low	Low	Medium
Low	Very High	Very High	Very low
Low	High	Low	Very low
Low	Medium	Low	Very Low
Low	Low	Very High	Medium

4.2.7 Phase 6: Risk Management

Encompass proposing and selecting appropriate strategies to mitigate identified risks. This phase includes an initial review of the output of phases 1-5 for human-centred validation of findings and initial planning of needed mitigations. On this basis, the risk management activities concern devising technical solutions like implementing robust data governance practices, access control mechanism, continuous model testing, encryption protocols, firewalls, and intrusion detection systems; It also involves human oversight measures such as training programs, co-creation workshops, policies, and procedures to enhance user awareness, trustworthy behaviour and technical skills. By carefully selecting countermeasures, organizations proactively strengthen the trustworthiness of their AI system. Effective countermeasures not only mitigate risks but also promote a culture of trustworthiness awareness and responsiveness throughout the organization.

Appendix A provides sources of controls that can be selected and implemented for every AI threat. The final decision for implementing controls is based on the criticality of the AI systems, risk appetite, a detailed cost benefit analysis, the business goals and strategies. Implementing controls to reduce risks, establishing continuous monitoring, and ensuring regulatory compliance are business decisions that the organisation that operates the AI system needs to undertake.

4.2.7.1 Output of Phase 6

A list of controls that have been selected, a list of controls that have been implemented, testing and evaluation reports of the controls.

A simple example that illustrates the six FAITH AI_TAF phases is included in Appendix D.

5 Implementation aspects of the FAITH AI_TAF

The FAITH AI_TAF will be implemented by the System Trust Modeler (STM). In particular, STM will support the second and third macro phases defined in Section 4, following an incremental approach.

The second stage is the characterization of the system under test by each LSP and risk calculation. By using the concepts defined above, the stakeholder performs a complete risk analysis following the process described in Section 4 and composed of 6 phases that reflect the ISO27005 process.

The last stage looks at the interaction of the AI system with other systems and the infrastructure in the LSP production environment where threats and consequences might be different and where again the ISO 27005 process is followed.

To support the characterization and risk calculation activities, WP2 provides a web-based application. The analysis focusses on the assessment of the AI system with its internal components during the different software lifecycle phases. The application will also include a form of decision support system providing pointers to relevant tools for the mitigation of the risks.

The STM will be described in detail in D2.3. Here, an initial sketch of the user journey is presented in Figure 14.

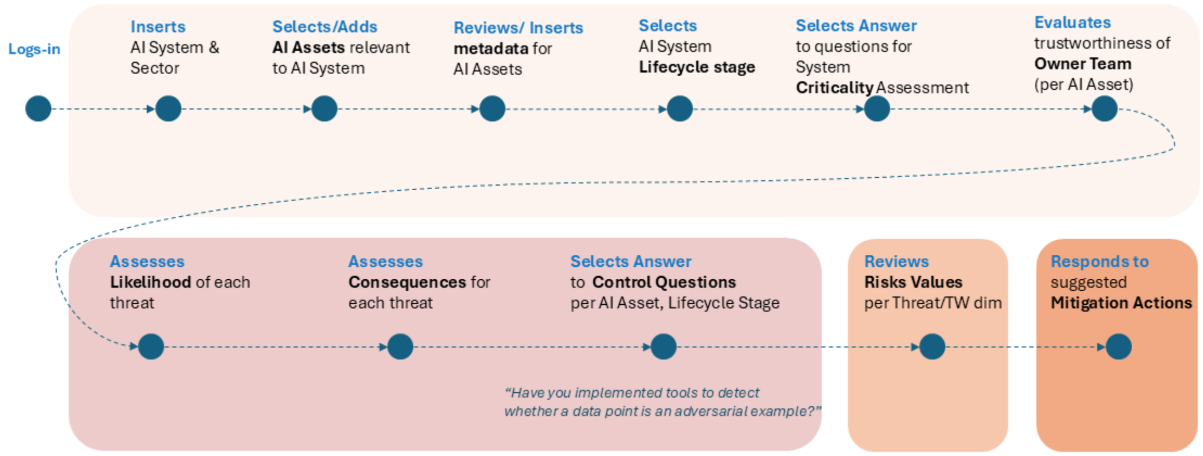

Logs-in
Inserts AI System & Sector
Selects/Adds AI Assets relevant to AI System
Reviews/ Inserts metadata for AI Assets
Selects AI System Lifecycle stage
Selects Answer to questions for System Criticality Assessment
Evaluates trustworthiness of Owner Team (per AI Asset)
Assesses Likelihood of each threat
Assesses Consequences for each threat
Selects Answer to Control Questions per AI Asset, Lifecycle Stage
"Have you implemented tools to detect whether a data point is an adversarial example?"
Reviews Risks Values per Threat/TW dim
Responds to suggested Mitigation Actions

Figure 13: User journey.

6 Preparation of Evaluation AI_TAF - v01

Here we develop the main principles that will be used for the evaluation methodology to assess the correctness and acceptability of the proposed framework (FAITH AI_TAF), assumptions, phases, measurements and its applicability for the various sectors (Domains).

6.1. Initial Evaluation Methodology

The evaluation methodology has been designed to ensure the needed rigor to faithfully assess the FAITH AI_TAF, while at the same time allow for its integration in the Faith TrustModeller (T2.2) tool and the necessary flexibility to fit the integration of the tool that range of FAITH pilots (T2.3).

The evaluation methodology needs to assess the implementation of the main FAITH AI_TAF main principles (Sees Section 4.1):

- ***Inclusive:*** Incorporates past experiences on risk assessment and builds upon them.
- ***Human centric:*** The trustworthiness of the AI participants is taken into account.
- ***Holistic:*** Applicable to each phase of the AI lifecycle.
- ***Tool independent****:* Independent of tools used.
- ***Agile:*** Can be used by all sectors (by adjusting the responses).
- ***Risk Assessment Standard based****:* Compliant with ISO27005 [23], ISO42001 (AI risk management) [145].
- ***Versatility:*** Beneficial to AI stakeholders regardless of the industry/sector.
- ***Global Reach*****:** Compliant with European and international initiatives, standards, and guidelines (e.g. ENISA FAICP, CEN/CENELEC, ETSI SAI, NIST AI RMF).

More than 10 meetings among the T2.1, T2.2. and T2.3 partners were conducted and more have been planned in order to align efforts during the WP2.

At the same time, the evaluation methodology will benefit from being applied by personnel that know well both the context and technology of the individual pilot. Furthermore, it will be beneficial to conduct evaluations in the regional or national language of the respective pilots.

For this purpose, the initial evaluation methodology can be laid out as a structure for evaluation workshops, for data collection from users, stakeholders and domain experts in the pilots. This methodology can then be implemented by research personnel from the respective pilots - as these know the pilot context, domain, and AI technology. This way, the workshops can also be held in the native language of the pilot users and stakeholders.

The evaluation workshops are a core component of the evaluation methodology. The workshops will include users, stakeholders and domain experts.

A very high-level workshop structure could, e.g., include the following components:

- Introduction to the FAITH AI_TAF/ specific components of the FAITH AI_TAF to be evaluated.
- Overview of evaluation aims and outcomes.
- Presentation and critical reflection on specific components of the FAITH AI_TAF. The presentation and reflection could take the form of a walkthrough. For each component, identification of positive aspects, negative aspects and potentials for improvement.
- Reflection on how the FAITH AI_TAF/ specific components may be implemented in the FAITH System Trust Modeller (T2.3)
- Measurements validation
- Demonstration of the tool that will implement the FAITH AI_TAF
- Operational Trial of the tool's functionality by the LSPs

The AI_TAF-V01 has already been presented in three (3) project meetings for comments. Based on these comments we provided this first version included in D2.1 .

1st FAITH Workshop: Introducing the concepts of trustworthiness in AI Systems and AI Participants"

The 1st physical workshop where all partners were invited was organised by trustilio and had the following objectives:

- Presentation of the AI_TAF phases
- Introduction of trustworthiness dimension and validation of measurements related to AI Systems and the Human Element (AI participants).
- Align our comprehension and our understanding of AI trustworthiness of AI Systems and the human involvement.
- Finalize measurements and scales for human maturity to respond to AI challenges and incidents
- **Workshop date/time:** 17 December / 10:00 am – 13:30 pm (GMT +3)

The validation of the Personality Traits and Organizational Maturity Scale for Technical Staff was conducted through a blended approach, combining virtual and in-person engagement with domain experts and project partners during the above workshop incorporating discussions, interactive sessions, and collaborative refinement of the scale. The validation process consisted of face and content validation, ensuring that the scale effectively measures the intended constructs and aligns with the broader goals of AI trustworthiness assessment. Following the workshop, participants were provided access to an online survey containing the draft scale. They were encouraged to test the survey and provide additional comments, ensuring that feedback extended beyond the workshop setting. The comments received were systematically reviewed and incorporated into the final version of the scale. Key areas of revision included:

⇒ Refining questions to focus on team-level assessment rather than individual attributes.

⇒ Introducing objective scaling mechanisms, such as defining frequency percentages for terms like "frequent failure," to enhance measurement consistency.

⇒ Balancing technical depth with accessibility, ensuring that both technical and non-technical users could effectively engage with the scale.

⇒ Aligning the questionnaire with organizational maturity models while maintaining a foundation in psychological assessment literature.

⇒ Standardizing the questionnaire for broader applicability, ensuring its relevance across various sectors where AI trustworthiness and data-driven decision-making are critical.

Based on the expert feedback and survey responses, the final submitted version of the scale was developed. The revisions incorporated adjustments in question wording, scoring mechanisms, while preserving the theoretical integrity of the measurement tool.

In the next iteration of the framework, partners will have the chance for further statistical analyses on pilot data to refine the scoring system and ensure its robustness in practical applications.

Similarly, internal meetings will reveal the enhancements for the next version of the FAITH_AI_TAF and the second workshop is planned to present the second version and validate the proposed measurements. The second version of the framework D.2.2 will finalise the measurements, uplift assumptions and further aligned with the functionality of the tool (D.2.3).

Similar workshops will be conducted in T2.2 and reported in D.2.3 and D.2.4.

7 Conclusions

The rapid advancement of Artificial Intelligence is reshaping socio-technical systems, aiming to replicate, augment, and, in some cases, replace human decision-making. AI engineering increasingly focuses on developing autonomous systems capable of analysing, predicting, and executing decisions with minimal human intervention. However, an overemphasis on AI system autonomy raises significant concerns in all requirements of trustworthiness (listed by HLEG on AI and referenced in the AI Act-see Appendix C) e.g.:

- Algorithmic bias and ethical inconsistencies that compromise fairness, equity
- Lack of transparency, reducing user confidence in AI-driven decisions.
- Diminished human oversight, leading to unintended consequences in critical applications.
- Increased cybersecurity vulnerabilities, where fully autonomous AI systems may inadvertently introduce new threats rather than mitigate existing risks.

While AI tools and techniques strive to enhance security and reliability, the absence of human integration in AI decision-making loops can leave organizations exposed to greater risks than anticipated. Addressing these risks necessitates a multi-disciplinary, trust-centric approach that aligns AI development with ethical, legal, social, and organizational principles and compliance with relevant legal instruments (e.g. AI Act, NIS2), standards and guidelines.

Trust in AI is not a singular concept, but a multi-dimensional construct shaped by several interdependent factors, including:

- Policy and legal frameworks governing AI accountability and compliance.
- Ethical and moral principles, ensuring AI systems align with societal values.
- Human perception and acceptance, influencing the extent to which AI is adopted in real-world applications.
- Technical performance and fairness, safeguarding AI reliability, accuracy, and robustness.
- Empowerment and transparency, fostering user confidence and enabling informed interactions with AI systems.

To enhance fairness and mitigate biases, organizations are leveraging tools such as IBM’s AI Fairness 360 and Google’s What-If Tool, which provide mechanisms to audit, interpret, and refine AI models. These solutions support the broader goal of AI trustworthiness by introducing methodologies that ensure AI behaves in ways that are predictable, explainable, and ethically sound.

To systematically assess and enhance AI trustworthiness, we propose the FAITH AI Trustworthiness Assessment Framework (FAITH_TAF). This framework integrates research from diverse disciplines—including AI governance, risk management, psychology, ethics, law, and cybersecurity—to establish a practical methodology for evaluating AI trustworthiness.

FAITH AI_TAF is designed to :

- Assess AI trustworthiness within its intended context, its life cycle, its purpose, criticality, usage, and the risk appetite of the organization.
- Evaluate an organization's AI maturity, identifying gaps in AI governance and proposing strategies for responsible AI adoption.
- Adopt a multilayer approach where Layer 1 provides the assumptions, Layer 2 estimates the general (by default) risk calculations and Layer 3 provides realistic estimates based on the environment, and maturity of teams. (Layers 2 and 3 are included in all phases of the FAITH AI_TAF).
- Extend and complement existing AI risk management frameworks, including:
  -NIST AI Risk Management Framework, which provides structured risk identification, assessment, and mitigation strategies.
  -ENISA AI Cybersecurity Framework, ensuring AI systems are secure, resilient, and aligned with European cybersecurity guidelines.
- Integrate socio-technical factors by accounting for human behaviour, social values, and business objectives throughout the AI risk management lifecycle.
- Emphasize co-creation and human experimentation, incorporating stakeholder feedback at every stage to continuously refine AI trustworthiness.
- Establish agile measurement methodologies, developing trustworthiness metrics and human-centric evaluation scales that align with both technical and societal expectations.

To validate the human trustworthiness scales integrated into FAITH_TAF, we conducted a FAITH workshop with AI domain experts, ethicists, cybersecurity professionals, and end-users. The workshop focused on:

- Understanding the socio-technical dynamics of AI trustworthiness.
- Refining trustworthiness metrics to reflect real-world AI applications.
- Introduce the TrustSense tool to measure AI maturity at human level.

The process, methodology, and outcomes of this initial FAITH workshop are detailed in the Appendix of this deliverable, providing empirical validation for the framework and setting the stage for further refinements.

The FAITH AI_TAF represents a significant advancement in AI governance, offering a structured, multi-dimensional approach to trustworthiness assessment. By integrating risk management principles, ethical AI considerations, and socio-technical factors, the FAITH AI_TAF ensures that AI systems:

- Remain transparent and accountable.
- Are designed with ethical safeguards in place.
- Align with human values and organizational responsibilities.
- Through co-creation, iterative evaluation, and stakeholder collaboration, the FAITH AI_TAF aims to establish AI ecosystems where users can confidently interact with AI technologies, fostering widespread acceptance and responsible AI innovation.

The FAITH AI_TAF will be provided in two (2) versions. In this version, presented in this deliverable, we made the assumptions:

- Data sources are assumed to be secure (they are certified/tested) (and not generate poisoned data).
- Each AI asset of the AI system under assessment has an owner (individual or team who is responsible for the asset).
- The responsibility for conducting a trustworthiness risk assessment process using the FAITH AI_TAF lies with the Risk Assessor or the ISMS coordinator of the organization (it is the organization's decision to select one person). All AI participants may be involved in the process, however in this 1st version of the FAITH AI_TAF only one person can provide the responses.
- We assume that no cascaded/propagated threats are feasible i.e., the AI system is either isolated or any interdependent components are secure (i.e., implemented controls do not enable the propagation of their vulnerabilities).

In the next version we will consider the uplift of the last assumption and also we will further validate the phases of the framework.

To implement the FAITH AI_TAF in practical AI governance settings, we propose a user journey-driven approach that informs the design and development of the FAITH Trust Modeler (a tool that will be developed under T2.2). This tool will:

- Assess and quantify AI trustworthiness across various dimensions, including ethical compliance, security, human perception, and organizational responsibility.
- Provide actionable insights for AI risk mitigation and policy alignment, ensuring AI systems adhere to best practices.
- Facilitate user interaction and feedback loops, enabling continuous refinement and adaptation of AI governance strategies.
- The FAITH Trust Modeler serves as a bridge between theoretical AI risk management principles and real-world application, ensuring AI solutions are trustworthy, transparent, and aligned with user expectations.

References

[1] Kaplan, A.D., et al. (2021). Trust in artificial intelligence: Meta-analytic findings. Human Factors: The Journal of the Human Factors and Ergonomics Society, 65(2), 337–359. doi:10.1177/00187208211013988.

[2] NIST - AI Risk Management Framework, (2024), https://www.nist.gov/itl/ai-risk-management-framework

[3] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Saxl, O. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines, 28(4), 689-707.

[4] Lipton, Z. C. (2016). The Mythos of Model Interpretability. In ICML Workshop on Human Interpretability in Machine Learning.

[5] Huang, Q., Li, Y., & Xiao, Y. (2020). Towards Dependable Artificial Intelligence: A Survey. Journal of Computer Science and Technology, 35(3), 487-515.

[6] Bryson, J. J., Diamantis, M. E., & Grant, T. D. (2017). Of, for, and by the people: The legal lacuna of synthetic persons. Artificial Intelligence and Law, 25(3), 273-291.

[7] Kioskli, K., Polemi, N., 'Estimating attackers' profiles results in more realistic vulnerability severity scores', in Ahram, T. and Karwowski, W. (eds), Human Factors in Cybersecurity, AHFE (2022) International Conference, AHFE Open Access, Vol. 53, AHFE International, 2022, http://doi.org/10.54941/ahfe1002211

[8] ENISA, Methodology for Sectoral Cybersecurity Assessments, 2021, https://www.enisa.europa.eu/publications/methodology-for-a-sectoral-cybersecurity-assessment.

[9] Akhavan Tabassi, Amin & Aldrin, Abdullah & Bryde, David. (2018). Conflict Management, Team Coordination, and Performance Within Multicultural Temporary Projects: Evidence From the Construction Industry. Project Management Journal. 50. 875697281881825. 10.1177/8756972818818257.

[10] Chakraborty, A., Alam, M., Dey, V., Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S., & Leung, V. C. M. (2018). A survey on security threats and defensive techniques of machine learning: A data-driven view. IEEE Access, 6, 12103-12117.

[11] Aufrant, L., and Hervieu, A. (2020). Guide de recommandations pour la spécification

et la qualification de systèmes intégrant de l'intelligence artificielle. Paris: DGA.

[12] Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. ACM SIGSAC Conference on Computer and Communications Security, 2154-2156.

[13] Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., & Kurakin, A. (2019). On evaluating adversarial robustness. arXiv:1902.06705.

[14] Barreno, M., Nelson, B., Joseph, A. D., & Tygar, J. D. (2010). The security of machine learning. Machine Learning, 81, 121-148.

[15] Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018). Sok: Security and privacy in machine learning. IEEE European Symposium on Security and Privacy (EuroS&P), 399-414.

[16] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In 2017 IEEE symposium on security and privacy (SP) (pp. 3-18). IEEE.

[17] ENISA, Multilayer Framework for Good Cybersecurity Practices for AI, 2023, https://www.enisa.europa.eu/publications/multilayer-framework-for-good-cybersecurity-practices-for-ai

[18] NIST, 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg: National Institute of Standards and Technology

[19] OECD, 2019. Recommendation of the Council on Artificial Intelligence. Paris: Organisation for Economic Co-operation and Development.

[20] OECD, 2024. Explanatory Memorandum on the Updated OECD Definition of an AI System. Paris: Organisation for Economic Co-operation and Development.

[21] ETSI, 2022. Securing Artificial Intelligence (SAI) AI Threat Ontology. Sophia Antipolis: European Telecommunications Standards Institute.

[22] ENISA, 2021. Securing Machine Learning Algorithms. Heraklion: European Union Agency for Cybersecurity

[23] ISO, n.d. ISO 2700x Standards. [online] Available at: https://www.iso.org/search.html?q=27000 [Accessed 8 July 2024]

[24] ISO/IEC, 2022. ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology. Geneva: International Organization for Standardization.

[25] ISO, 2018. ISO 31000:2018 Risk management — Guidelines. Geneva: International Organization for Standardization

[26] ISO, 2015. ISO 9000:2015 Quality management systems — Fundamentals and vocabulary. Geneva: International Organization for Standardization.

[27] ISO, 2022. ISO/IEC TS 5723:2022 Trustworthiness — Vocabulary. Geneva: International Organization for Standardization

[28] European Commission, Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment, https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

[29] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>

[30] See further Donatella Casaburo and Lorenzo Gugliotta, 'The EU AI Act proposal(s): Manipulative and exploitative AI practices', CiTiP Blog (22 September 2023), available at https://www.law.kuleuven.be/citip/blog/the-eu-ai-act-proposals-manipulative-and-exploitative-ai-practices/.

[31] Voss, W. Gregory, AI Act: The European Union's proposed framework regulation for Artificial Intelligence Governance, 25 Journal of Internet Law 4 (2021), at 10.

[32] DIRECTIVE (EU) 2022/2555 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive) <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022L2555>

[33] Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013 (Cybersecurity Act) <https://eur-lex.europa.eu/eli/reg/2019/881/oj>

[34] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on horizontal cybersecurity requirements for products with digital elements and amending Regulation (EU) 2019/1020 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52022PC0454

[35] OECD - Artificial Intelligence, 2019, https://www.oecd.org/digital/artificial-intelligence/

[36] H. JamesWilson and Paul R. Daugherty. 2018. Collaborative intelligence: Humans and AI are joining forces. Harvard Business Review 96, 4 (2018), 114–123.

[37] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, andDawn Song. 2017. Robust physical-world attacks on deep learning models. arXiv preprint arXiv:1707.08945 (2017).

[38] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, and Prateek Mittal. 2018. Darts: Deceiving autonomous cars with toxic signs. arXiv preprint arXiv:1802.06430 (2018).

[39] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. 2017. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. arXiv:1712.05526

[40] Qi Lei, Lingfei Wu, Pin-Yu Chen, Alexandros G. Dimakis, Inderjit S. Dhillon, and Michael Witbrock. 2018. Discrete attacks and submodular optimization with applications to text classification. CoRR abs/1812.00151 (2018). arXiv:1812.00151 http://arxiv.org/abs/1812.00151.

[41] Haochen Liu, Tyler Derr, Zitao Liu, and Jiliang Tang. 2019. Say what I want: Towards the dark side of neural dialogue models. arXiv preprint arXiv:1909.06044 (2019).

[42] Nicholas Carlini and David Wagner. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In 2018 IEEE Security and Privacy Workshops (SPW). IEEE, 1–7.

[43] Nicolae, Maria-Irina, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, et al. “Adversarial Robustness Toolbox v1.0.0.” arXiv, November 15, 2019.

[44] Papernot, Nicolas, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, et al. “Technical Report on the CleverHans v2.1.0 Adversarial Examples Library.” arXiv, June 27, 2018.

[45] Li, Yaxin, Wei Jin, Han Xu, and Jiliang Tang. “DeepRobust: A PyTorch Library for Adversarial Attacks and Defenses.” arXiv, May 13, 2020.

[46] Croce, Francesco, et al. "Robustbench: a standardized adversarial robustness benchmark." arXiv preprint arXiv:2010.09670 (2020).

[47] Goodman, Dou, Hao Xin, Wang Yang, Wu Yuesheng, Xiong Junfeng, and Zhang Huan. “Advbox: A Toolbox to Generate Adversarial Examples That Fool Neural Networks.” arXiv, August 26, 2020.

[48] Ding, Gavin Weiguang, Luyu Wang, and Xiaomeng Jin. “Advertorch v0.1: An Adversarial Robustness Toolbox Based on PyTorch.” arXiv, February 20, 2019.

[49] Rauber, Jonas, Wieland Brendel, and Matthias Bethge. “Foolbox: A Python Toolbox to Benchmark the Robustness of Machine Learning Models.” arXiv, March 20, 2018.

[50] “Giskard - Testing Platform for AI Systems.” https://www.giskard.ai/.

[51] Morris, John X., Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. “TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP.” arXiv, October 5, 2020.

[52] Kim, Hoki. “Torchattacks: A PyTorch Repository for Adversarial Attacks.” arXiv, February 19, 2021.

[53] ISO, 2017. ISO/IEC 27001:2017 Information technology — Security techniques — Information security management systems — Requirements. Geneva: International Organization for Standardization. https://www.iso.org/standard/45481.html

[54] ISO, 2021. ISO/IEC 42001:2021 Artificial intelligence — Management systems — Requirements with guidance for use. Geneva: International Organization for Standardization. https://www.iso.org/standard/81608.html

[55] Pintz, Maximilian, Joachim Sicking, Maximilian Poretschkin, and Maram Akila. “A Survey on Uncertainty Toolkits for Deep Learning.” arXiv, May 2, 2022.

[56] Dürr, Oliver, Beate Sick, and Elvis Murina. Probabilistic deep learning: With python, keras and tensorflow probability. Manning Publications, 2020.

[57] Bingham, Eli, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. “Pyro: Deep Universal Probabilistic Programming.” arXiv, October 18, 2018.

[58] Ghosh, Soumya, Q. Vera Liao, Karthikeyan Natesan Ramamurthy, Jiri Navratil, Prasanna Sattigeri, Kush R. Varshney, and Yunfeng Zhang. “Uncertainty Quantification 360: A Holistic Toolkit for Quantifying and Communicating the Uncertainty of AI.” arXiv, June 4, 2021.

[59] Gardner, Jacob R., Geoff Pleiss, David Bindel, Kilian Q. Weinberger, and Andrew Gordon Wilson. “GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration.” arXiv, June 29, 2021.

[60] Chung, Youngseog, Ian Char, Han Guo, Jeff Schneider, and Willie Neiswanger. “Uncertainty Toolbox: An Open-Source Library for Assessing, Visualizing, and Improving Uncertainty Quantification.” arXiv, September 21, 2021.

[61] Duan, Tony, Avati Anand, Daisy Yi Ding, Khanh K. Thai, Sanjay Basu, Andrew Ng, and Alejandro Schuler. “NGBoost: Natural Gradient Boosting for Probabilistic Prediction.” In Proceedings of the 37th International Conference on Machine Learning, 2690–2700. PMLR, 2020.

[62] Amazon Recommender System, Advisors: Ilkay Altintas, Julian McAuley Team Members: JH (Janghyun) Baek, John Tsai, Justin Shamoun, Muriel Marable, Ying Cui, https://library.ucsd.edu/dc/object/bb8503744c/_2_1.pdf

[63] Ma, W., Zhang, M., Cao, Y., Jin, W., Wang, C., Liu, Y., ... & Ren, X. (2019, May). Jointly learning explainable rules for recommendation with knowledge graph. In The world wide web conference (pp. 1210-1221).

[64] Vig, Jesse. “A Multiscale Visualization of Attention in the Transformer Model.” arXiv, June 12, 2019.

[65] “Captum · Model Interpretability for PyTorch.” https://captum.ai/.

[66] Wang, Zijie J., Robert Turko, Omar Shaikh, Haekyu Park, Nilaksh Das, Fred Hohman, Minsuk Kahng, and Duen Horng Polo Chau. “CNN Explainer: Learning Convolutional Neural Networks with Interactive Visualization.” IEEE Transactions on Visualization and Computer Graphics 27, no. 2 (February 2021): 1396–1406.

[67] Yosinski, Jason, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. “Understanding Neural Networks Through Deep Visualization.” arXiv, June 22, 2015.

[68] Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." arXiv, December 3, 2019.

[69] Nori, Harsha, Samuel Jenkins, Paul Koch, and Rich Caruana. "InterpretML: A Unified Framework for Machine Learning Interpretability." arXiv, September 19, 2019.

[70] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." arXiv, August 9, 2016.

[71] "Netron." https://netron.app/.

[72] "Utkuozbulak/Pytorch-Cnn-Visualizations: Pytorch Implementation of Convolutional Neural Network Visualization Techniques." https://github.com/utkuozbulak/pytorch-cnn-visualizations.

[73] Lundberg, Scott, and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." arXiv, November 25, 2017.

[74] Arya, Vijay, Rachel K E Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, et al. "AI Explainability 360: An Extensible Toolkit for Understanding Data and Machine Learning Models".

[75] Liu, M., Luo, Y., Wang, L., Xie, Y., Yuan, H., Gui, S., ... & Ji, S. (2021). DIG: A turnkey library for diving into graph deep learning research. Journal of Machine Learning Research, 22(240), 1-9.

[76] Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2017). Towards better understanding of gradient-based attribution methods for deep neural networks. arXiv preprint arXiv:1711.06104.

[77] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradientbased attribution methods for deep neural networks. In International Conference on Learning Representations.

[78] Georgios A. Kaissis, Marcus R. Makowski, Daniel Rückert, and Rickmer F. Braren. 2020. Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence 2, 6 (2020), 305–311.

[79] Fida Kamal Dankar and Khaled El Emam. 2013. Practicing differential privacy in health care: A review. Trans. Data Priv. 6, 1 (2013), 35–67.

[80] Esma Aïmeur, Gilles Brassard, José M. Fernandez, and Flavien Serge Mani Onana. 2008. A lambic: A privacypreserving recommender system for electronic commerce. International Journal of Information Security 7, 5 (2008), 307–334.

[81] Holohan, Naoise, Stefano Braghin, Pól Mac Aonghusa, and Killian Levacher. “Diffprivlib: The IBM Differential Privacy Library.” arXiv, July 4, 2019.

[82] Liu, Yang, Tao Fan, Tianjian Chen, Qian Xu, and Qiang Yang. “FATE: An Industrial Grade Platform for Collaborative Learning With Data Protection.” Journal of Machine Learning Research 22, no. 226 (2021): 1–6.

[83] He, Chaoyang, Songze Li, Jinhyun So, Xiao Zeng, Mi Zhang, Hongyi Wang, Xiaoyang Wang, et al. “FedML: A Research Library and Benchmark for Federated Machine Learning.” arXiv, November 8, 2020.

[84] Beutel, Daniel J., Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, et al. “Flower: A Friendly Federated Learning Research Framework.” arXiv, March 5, 2022.

[85] Halevi, Shai, and Victor Shoup. “Design and Implementation of HElib: A Homomorphic Encryption Library,” 2020. Cryptology ePrint Archive. https://eprint.iacr.org/2020/1481.

[86] “Open Policy Agent.” https://www.openpolicyagent.org/.

[87] Ziller, Alexander, Andrew Trask, Antonio Lopardo, Benjamin Szymkow, Bobby Wagner, Emma Bluemke, Jean-Mickael Nounahon, et al. “PySyft: A Library for Easy Federated Learning.” In Federated Learning Systems: Towards Next-Generation AI, edited by Muhammad Habib ur Rehman and Mohamed Medhat Gaber, 111–39. Cham: Springer International Publishing, 2021.

[88] TensorFlow. “TensorFlow Privacy | Responsible AI Toolkit.” https://www.tensorflow.org/responsible_ai/privacy/guide.

[89] TensorFlow. “TensorFlow Federated.” https://www.tensorflow.org/federated.

[90] FAITH-FORTH, 2023. DPA - Data Protection Assessment. Available at: https://github.com/FAITH-FORTH/DPA [Accessed 9 February 2025].

[91] 2008. IEEE Standard for Software Reviews and Audits. IEEE Std 1028-2008 (2008), 1–53. https://doi.org/10.1109/IEEESTD.2008.4601584

[92] Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. “Model Cards for Model Reporting.” In Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–29. Atlanta, GA USA: ACM, 2019.

[93] Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. “Datasheets for Datasets.” arXiv, December 1, 2021.

[94] Arnold, Matthew, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, et al. "FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity." arXiv, February 7, 2019.

[95] Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. "Model Cards for Model Reporting." In Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–29. Atlanta, GA USA: ACM, 2019.

[96] Ayanna Howard and Jason Borenstein. 2018. The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. Science and Engineering Ethics 24, 5 (2018), 1521–1536.

[97] James A. Rodger and Parag C. Pendharkar. 2004. A field study of the consequence of gender and user's technical experience on the performance of voice-activated medical tracking application. International Journal of Human-Computer Studies 60, 5-6 (2004), 529–544.

[98] NIST, 2023. AI Risk Management Framework. Available at: https://www.nist.gov/itl/ai-risk-management-framework [Accessed 9 February 2025].

[99] NIST, 2023. NIST Special Publication 1270: A Cybersecurity Framework for Artificial Intelligence. Available at: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf [Accessed 9 February 2025].

[100] ISO, 2023. ISO/IEC 20546:2023 Information technology — Big data — Reference architecture. Geneva: International Organization for Standardization. Available at: https://www.iso.org/standard/77607.html [Accessed 9 February 2025].

[101] Bellamy, Rachel K. E., Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, et al., "AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias." arXiv, October 3, 2018.

[102] Weerts, Hilde, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. "Fairlearn: Assessing and Improving Fairness of AI Systems." arXiv, March 29, 2023.

[103] Bantilan, Niels. "Themis-Ml: A Fairness-Aware Machine Learning Interface for End-to-End Discrimination Discovery and Mitigation." arXiv, October 18, 2017.

[104] Baumann, Joachim, Alessandro Castelnovo, Riccardo Crupi, Nicole Inverardi, and Daniele Regoli. "Bias on Demand: A Modelling Framework That Generates Synthetic Data With Bias." In 2023 ACM Conference on Fairness, Accountability, and Transparency, 1002–13. Chicago IL USA: ACM, 2023.

[105] Saleiro, Pedro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. "Aequitas: A Bias and Fairness Audit Toolkit." arXiv, April 29, 2019.

[106] Zehlike, Meike, Carlos Castillo, Francesco Bonchi, Ricardo Baeza-Yates, Sara Hajian, Mohamed Megahed. "Fairness Measures: Datasets and software for detecting algorithmic discrimination." June, 2017. http://fairness-measures.org/.

[107] Sokol, Kacper, Raul Santos-Rodriguez, and Peter Flach. "FAT Forensics: A Python Toolbox for Algorithmic Fairness, Accountability and Transparency." Software Consequences 14 (December 1, 2022): 100406.

[108] Wang, Angelina, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. "REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets." arXiv, July 23, 2021.

[109] FAITH-FORTH, 2023. DBDM - Data-Based Decision Making. Available at: https://github.com/FAITH-FORTH/DBDM [Accessed 9 February 2025].

[110] FAITH-FORTH, 2023. MANDALA - AI-Based Decision Making Framework. Available at: https://github.com/FAITH-FORTH/MANDALA [Accessed 9 February 2025].

[111] Pagano, Tiago P., et al. "Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods." Big data and cognitive computing 7.1 (2023): 15.

[112] Hardt, Michaela, et al. "Amazon sagemaker clarify: Machine learning bias detection and explainability in the cloud." Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021.

[113] Tramer, Florian, et al. "Fairtest: Discovering unwarranted associations in data-driven applications." 2017 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, 2017.

[114] OECD, 2023. AI Metrics. Available at: https://oecd.ai/en/catalogue/metrics [Accessed 9 February 2025].

[115] Templ, Matthias. "Statistical disclosure control for microdata." Cham: Springer (2017).

[116] Vakili, Meysam, Mohammad Ghamsari, and Masoumeh Rezaei. "Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification." arXiv, January 27, 2020.

[117] Google for Developers. "Classification: ROC and AUC | Machine Learning."

[118] "Mean Squared Error." In Wikipedia, June 11, 2024.

[119] "Dice-Sørensen Coefficient." In Wikipedia, December 10, 2024.

[120] Rezatofighi, Hamid, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. "Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression." In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 658–66. Long Beach, CA, USA: IEEE, 2019.

[121] Celaya, Adrian, Beatrice Riviere, and David Fuentes. "A Generalized Surface Loss for Reducing the Hausdorff Distance in Medical Imaging Segmentation." arXiv, January 24, 2024.

[122] Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "Bleu: A Method for Automatic Evaluation of Machine Translation." In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, edited by Pierre Isabelle, Eugene Charniak, and Dekang Lin, 311–18. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002.

[123] Lin, Chin-Yew. "ROUGE: A Package for Automatic Evaluation of Summaries." In Text Summarization Branches Out, 74–81. Barcelona, Spain: Association for Computational Linguistics, 2004.

[124] Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. "BERTScore: Evaluating Text Generation with BERT." arXiv, February 24, 2020.

[125] Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. "On Calibration of Modern Neural Networks." arXiv, August 3, 2017.

[126] Sun, Yiyou, Yifei Ming, Xiaojin Zhu, and Yixuan Li. "Out-of-Distribution Detection with Deep Nearest Neighbors." In Proceedings of the 39th International Conference on Machine Learning, 20827–40. PMLR, 2022.

[127] Walter, S. (2022). User Journey Mapping. Publisher(s): SitePoint, ISBN: 9781925836493

[128] UCI Information Security Standard, 2024, https://security.uci.edu/security-plan/plan-controls.html

[129] Parsa Saadatpanah, Ali Shafahi, and Tom Goldstein. 2020. Adversarial attacks on copyright detection systems. In International Conference on Machine Learning. PMLR, 8307–8315.

[130] Gavin Weiguang Ding, Luyu Wang, and Xiaomeng Jin. 2019. AdverTorch v0.1: An adversarial robustness toolbox based on Pytorch. arXiv preprint arXiv:1902.07623 (2019).

[131] Yaxin Li, Wei Jin, Han Xu, and Jiliang Tang. 2020. DeepRobust: A PyTorch Library for Adversarial Attacks and Defenses. arXiv:2005.06149

[132] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. 2021. RobustBench: A Standardized Adversarial Robustness Benchmark. arXiv:2010.09670

[133] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on Fairness, Accountability and Transparency. 77–91.

[134] Carty S. 2011. Many cars tone deaf to women's voices. AOL Autos (2011).

[135] Wenqi Fan, Tyler Derr, Xiangyu Zhao, Yao Ma, Hui Liu, Jianping Wang, Jiliang Tang, and Qing Li. 2021. Attacking black-box recommendations via copying cross-domain user profiles. In 2021 IEEE 37th International Conference on Data Engineering (ICDE). IEEE, 1583–1594.

[136] Weizhi Ma, Min Zhang, Yue Cao, Woojeong Jin, Chenyang Wang, Yiqun Liu, Shaoping Ma, and Xiang Ren. 2019. Jointly learning explainable rules for recommendation with knowledge graph. In The World Wide Web Conference. 1210–1221.

[137] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilovic, et al. 2020. AI explainability 360: An extensible toolkit for understanding data and machine learning models. Journal of Machine Learning Research 21, 130 (2020), 1–6.

[138] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. InterpretML: A unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223 (2019).

[139] Meng Liu, Youzhi Luo, Limei Wang, Yaochen Xie, Hao Yuan, Shurui Gui, Haiyang Yu, Zhao Xu, Jingtun Zhang, Yi Liu, et al. 2021. DIG: A turnkey library for diving into graph deep learning research. arXiv preprint arXiv:2103.12608 (2021).

[140] Michael Kapralov and Kunal Talwar. 2013. On differentially private low rank approximation. In Proceedings of the Twenty-fourth Annual ACM-SIAM Symposium on Discrete Algorithms. SIAM, 1395–1414.

[141] Nicholas Carlini, Florian Tramer, EricWallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In 30th USENIX Security Symposium (USENIX Security 21). 2633–2650.

[142] Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. 2021. Large-Scale Differentially Private BERT. arXiv:2108.01624

[143] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In Proceedings of the Conference on Fairness, Accountability, and Transparency. 90–99.

[144] ENISA, AI Cybersecurity Challenges, Threat landscape for Artificial intelligence, 2020, https://www.enisa.europa.eu/sites/default/files/publications/ENISA%20Report%20-%20Artificial%20Intelligence%20Cybersecurity%20Challenges.pdf

[145] ISO/IEC, 2023. ISO/IEC 42001:2023 Artificial Intelligence — Management Systems — Requirements with Guidance for Use. Geneva: International Organization for Standardization.

Appendices

Appendix A: FAITH Knowledge Base

The current version of knowledge base (KB) of FAITH, including the Golden Table produced by FORTH, provides a comprehensive reference for relevant data and developments. This annex presents these details for further review and analysis. Additionally, it includes insights, covering AI lifecycle stages as defined by ENISA and FAITH, along with key aspects such as security, resilience, explainability, privacy enhancement, fairness, accountability, and reliability. The annex also outlines identified threats—including evasion and poisoning attacks—linked vulnerabilities, and corresponding controls, providing a structured approach to AI risk management. Further details are included in the provided link.

Appendix B: 1st FAITH Workshop: Introducing the concepts of trustworthiness in AI Systems and AI Participants

- **Workshop Objectives:**
  - Introduction of trustworthiness dimension and validation of measurements related to AI Systems and the Human Element (AI participants).
  - Align our comprehension and our understanding of AI trustworthiness of AI Systems and the human involvement.

- **Workshop date/time:** 17 December / 10:00 am – 13:30 pm (GMT +3)
- **Workshop Venue:** Hybrid

Agenda	
Time (GMT +3)	**Topic**
10:00 – 10:30	Introducing FAITH Trustworthiness Framework
10:30 – 11:00	AI Threats, Vulnerabilities and Controls (NewRisk.xlsx)
11:00 – 11:30	Human AI Trustworthiness and Oversight (AI Trustworthiness dimensions / Experiences, Preferences and Priorities / AI Actors / Participants)
11:30 – 12:00	AI Participants Trustworthiness Measurement
12:00 – 12:15	Coffee Break
12:15 – 13:15	Measurement Testing / Proposed FAITH Scales / Scales Consensus
13:15 – 13:30	Measurement Consensus Building / Conclusions

Appendix C: Glossary

References: ISO22989, ISO/IEC 27000:2018, ISO 31000:2018, ISO/IEC 27042:2015, JRC Glossary of human-centric artificial intelligence, ChatGPTv4

A

- **AI (Artificial Intelligence):** The simulation of human intelligence in machines, designed to perform tasks like learning, reasoning, and problem-solving.
- **Accountability** means having clear ownership and responsibility for the outcomes of AI systems. **Accountable**: answerable for actions, decisions and performance.
- **Attacker/Adversary**: actor that potentially uses a vulnerability (weakness) to exploit a threat(s).
- **Attack potential**: measure of the effort needed to exploit a vulnerability in a target.
- **Availability**: Property of being accessible and usable on demand by an authorised entity.

B

- **Bias**: Inclination of prejudice towards or against a person, object, or position. Bias can arise in many ways in AI systems. For example, in data-driven AI systems, such as those produced through machine learning, bias in data collection and training can result in an AI system demonstrating bias. In logic-based AI, such as rule-based systems, bias can arise due to how a knowledge engineer might view the rules that apply in a particular setting. Bias can also arise due to online learning and adaptation through interaction. It can also arise through personalisation whereby users are presented with or human-driven data collection. It can arise, for example, through the limited contexts in which a system is used, in which case there is no opportunity to generalise it to other contexts. Bias can be good or bad, intentional or unintentional. In certain cases, bias can result in discriminatory and/or unfair outcomes, indicated in (HLEG AI, 2019) as unfair bias.
- **Bias (Algorithmic):** systematic difference in treatment of certain objects, people or groups in comparison to others. A systematic error in AI models that leads to unfair or incorrect outcomes, often due to imbalanced training data.
- **Bias Mitigation** involves identifying, addressing, and reducing biases that may exist in data or algorithms. Biased data or models can result in unfair treatment based on race, gender, age, or other characteristics. Mitigation activities include: Data balancing, fairness constraints, and diverse datasets help reduce these biases.

C

- **Controls:** Controls are defined as policies, procedures, or technical measures put in place to manage, monitor, and regulate the performance of systems, processes, or activities. In the context of AI, controls are designed to ensure that AI systems operate in a trustworthy manner, comply with relevant regulations, and achieve desired outcomes while mitigating risks to users and society.

- **Cybersecurity:** An AI system ensures the confidentiality, integrity, authenticity and availability of all its components (networks, data, algorithms, models, processes, users/participants) at all stages of its lifecycle.
- **Chatbot:** An AI-driven program designed to simulate conversation with human users, often used in customer service and support.
- **Classification:** A machine learning task of categorizing data into predefined labels or classes.
- **Clustering:** A technique used in unsupervised learning to group data points that are similar to each other into clusters.

D

- **Deep Learning:** A subset of machine learning involving neural networks with many layers, designed to model complex patterns in large datasets.
- **Data augmentation**: process of creating synthetic samples by modifying or utilizing the existing data.
- **Data sampling**: process to select a subset of data samples intended to present patterns and trends similar to that of the larger dataset (3.2.5) being analysed.
- **Dataset**: collection of data with a shared format.

E

- **Explainable AI** (XAI) methods, such as LIME or SHAP, provide insights into how decisions are made by black-box models like neural networks.
- **Ethics in AI:** The consideration of moral implications and potential biases in AI systems, ensuring that AI is developed responsibly.
- **Ethical AI**: term used to indicate the development, deployment and use of AI that ensures compliance with ethical norms, including fundamental rights as special moral entitlements, ethical principles, and related core values.

F

- **Fairness** refers to a variety of ideas known as equity, impartiality, egalitarianism, non-discrimination and justice. Fairness embodies an ideal of equal treatment between individuals or between groups of individuals. This is what is generally passes a procedural perspective, that is the ability to seek and obtain relief when individual rights and freedoms are violated.
- **Feature Extraction:** The process of identifying key characteristics or attributes from raw data that can be used for model building.

G

- **Generative AI (GenAI):** type of AI system that addresses a broad range of tasks with a satisfactory level of performance. GenAI creates new data or content, such as images, text, or music, based on learned patterns.

- **Generative Adversarial Network (GAN):** A class of machine learning frameworks where two neural networks (a generator and a discriminator) compete to generate realistic data.
- **Governance of AI**: Establishing frameworks and guidelines that outline responsible use of AI in organizations.

H

- **Heuristic:** A rule of thumb or strategy for problem-solving that is not guaranteed to be optimal but can yield quick, satisfactory results.
- **Human-Centric Design**: AI systems designed to work alongside humans, enhancing human decision-making rather than replacing it. This includes creating user-friendly interfaces and maintaining a level of human oversight to prevent unintended consequences.
- **Human-in-the-loop (HITL)**: humans are involved in critical decision-making processes or in monitoring AI outcomes for accountability; it is one of the governance mechanisms addressed by human oversight. HITL refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable.

I

- **Image Recognition:** The ability of AI systems to interpret and classify objects within an image.
- **Instance:** An individual example or data point in a dataset used for training or testing machine learning models.

- **Information Security Management System (ISMS):** a systematic approach to managing company information, so that it remains secure. It includes people, processes, and IT systems by applying a risk management approach.

K

- **K-Means Clustering:** A simple and commonly used algorithm for dividing a dataset into clusters by finding groups based on their distance from a central point.
- **Knowledge Representation:** How information and rules are structured for AI to interpret and reason over, such as logic or semantic networks.

L

- **Language Model:** An AI model designed to predict the likelihood of sequences of words, enabling tasks like translation, summarization, or text generation.
- **Latent Space:** A lower-dimensional representation of data learned by models, such as autoencoders or GANs, to capture important features.

- **Lifecycle of an AI system:** evolution of a system, product, service, project or other human-made entity, from conception through retirement; design, development, tastings and production phases.

M

- **Machine Learning (ML):** Process of optimizing model parameters through computational techniques, such that the model's behaviour reflects the data or experience. A subset of AI where algorithms improve their performance at a task through experience, without explicit programming.
- **Machine Learning Platforms:** Provide an ecosystem of tools, libraries and resources that support the development of machine learning applications.
- **Mitigation Actions:** Mitigation actions refer to specific measures or strategies implemented to reduce or eliminate identified risks or negative consequences associated with a system, process, or activity. In the context of AI, these actions aim to address potential trustworthiness concerns, such as fairness, security, or transparency, by implementing proactive steps to minimize harm and enhance system reliability.
- **Model:** physical, mathematical or otherwise logical representation of a system, entity, phenomenon, process or data. A mathematical representation of a real-world process or system created by machine learning algorithms from training data.
- **Meta-learning:** Also known as “learning to learn,” it refers to AI algorithms that learn to improve their learning process over time.
- **Multimodal Learning:** AI models that can process and integrate data from different modalities, like text, images, and sound.
- **Model poisoning:** In the context of AI as a Service, with many types of data and code being uploaded on cloud infrastructures, this threat may be realized by exploiting potential weaknesses of cloud providers.

N

- **Natural Language Processing (NLP):** The field of AI that focuses on the interaction between computers and humans through natural language.
- **Neural Network:** A computational model inspired by the human brain, consisting of layers of nodes (neurons) that learn to perform tasks by adjusting weights.
- **Normalization:** A preprocessing step in machine learning where data is adjusted to have a standard scale, improving algorithm performance.

O

- **Optimization:** The process of improving a machine learning model by finding the best parameters that minimize the loss function.

P

- **AI Participant:** AI users that are involved/participate/interact at the various phases of the AI systems' lifecycle.
- **Personal data:** any information that (a) can be used to establish a link between the information and the natural person to whom such information relates, or (b) is or can be directly or indirectly linked to a natural person
- **Prompt Engineering:** The practice of carefully designing and refining input prompts to get desired responses from language models or generative AI.
- **Predictability**: property of an AI system that enables reliable assumptions by stakeholders about the output.
- **Predictive Analytics:** The use of AI to analyze current and historical data to make predictions about future outcomes.
- **Pretraining:** The initial phase of training a model on a large dataset before fine-tuning it for a specific task.

Q

- **Quality:** process in which data is examined for completeness, bias and other factors which affect its usefulness for an AI system.

R

- **Reliability**: property of consistent intended behaviour and results.
- **Resilience ability** of a system to recover operational condition quickly following an incident.
- **Responsibility**: Capability of fulfilling an obligation or duty; The quality of being reliable or trustworthy; the state or fact of being accountable for actions; liability for some action.
- **Risk** is expressed in terms of threats, vulnerabilities and consequences.
- **Robustness** of a system to maintain its level of performance under any circumstances refers to the resilience of AI systems to handle unexpected inputs, attacks, or changes in the environment without failure.
- **Reinforcement Learning (RL):** A machine learning paradigm where agents learn to make decisions by receiving rewards or penalties from their environment based on actions.
- **Recurrent Neural Network (RNN):** A type of neural network designed for sequential data, where connections between nodes form cycles to maintain memory of previous inputs.
- **Resilience:** The ability of the AI system to respond and recover from attacks, failures, unexpected disruptions.

S

- **Safety** (physical security) ensures that AI systems operate in ways that avoid causing harm to humans, infrastructure, or the environment.
- **Supervised Learning:** A type of machine learning where the model is trained on labelled data, meaning each input has an associated correct output.
- **Support Vector Machine (SVM):** A supervised learning algorithm used for classification and regression tasks by finding the optimal separating boundary between classes.

T

- **Training data**: data used to train a machine learning model
- **Trustworthiness** refers to the confidence that users, organizations, and society can place in artificial intelligence systems to behave in a reliable, safe, ethical, and transparent manner.
- **Trustworthiness requirements:** AI Act refers to the seven key requirements for trustworthy AI defined by the EU's High-Level Expert Group on Artificial Intelligence (AI HLEG). These requirements are:
- **Human Agency and Oversight:** Ensuring AI systems support human autonomy and decision-making, with appropriate human oversight mechanisms.
- **Technical Robustness and Safety:** Guaranteeing that AI systems are resilient, secure, and reliable, minimizing potential risks and errors.
- **Privacy and Data Governance:** Protecting personal data and ensuring its proper management and use within AI systems.
- **Transparency:** Providing clear information about AI system operations, capabilities, and limitations to foster understanding and trust.
- **Diversity, Non-discrimination, and Fairness:** Preventing bias and ensuring equitable treatment of all individuals and groups by AI systems.
- **Societal and Environmental Well-being:** Promoting positive social and environmental impacts through the use of AI.
- **Accountability:** Establishing mechanisms to ensure responsibility and accountability for AI system outcomes
- **Trustworthiness dimensions**: FAITH trustworthiness dimensions considered are:
- Safety
- Security and Resilience
- Explainability and Interpretability
- Privacy
- Fairness, Non Bias
- Accountability and transparency
- Validity and Reliability
- **(security) threat:** potential cause (intentional or unintentional) of an information security incident which may result in harm to the security (confidentiality, integrity, authenticity) of an AI system.
- **(trustworthiness) threat**: potential cause (intentional or unintentional) of an incident which may result in harm to the trustworthiness of an AI system.

- **Transparency**: property of an organization that appropriate activities and decisions are communicated to relevant stakeholders in a comprehensive, accessible and understandable manner. Appropriate information for system transparency can include aspects such as features, performance, limitations, components, procedures, measures, design goals, design choices and assumptions, data sources and labelling protocols. Inappropriate disclosure of some aspects of a system can violate security, privacy or confidentiality requirements. refers to how openly the AI system's processes, decisions, and functionalities are communicated to users and stakeholders.
- **Trustworthy AI participant**: An AI participant with human traits and characteristics (e.g., socio-psychological, behavioural, capabilities, skills) that can ensure the trustworthiness of the AI system.
- **Transfer Learning:** The process of leveraging a pre-trained model on one task to apply it to another related task, reducing training time and improving performance.

U

- **Unsupervised Learning:** A type of machine learning where the model is trained on data without labelled outputs, used for discovering hidden patterns or structures in the data.

V

- **Validation data** used to compare the performance of different candidate models
- **Variational Autoencoder (VAE):** A type of generative model that learns the probability distribution of data and can generate new data points.
- **Vectorization:** The process of converting data (e.g., text) into a numerical form (vectors) that can be processed by machine learning algorithms.
- **Vulnerability**: weakness of an asset or control that can be exploited by one or more threats.

W

- **Workflow of the model**: The workflow of an AI model shows the phases needed to build the model and their interdependencies. Typical phases are: Model usage, Model maintenance, Model versioning. These stages are usually iterative: one may need to re-evaluate and go back to a previous step at any point in the process.

X

Y

Z

- **Zero day vulnerability:** refers to a security flaw or software weakness that is unknown to the vendor or developer and has not yet been patched or fixed.

Appendix D: Example for illustrating the AI-TAF

In this example, we selected as AI system under assessment**:** a simple AI-based chatbot designed to provide customer support for an e-commerce platform. This chatbot responds to customer queries, processes orders, and provides troubleshooting guidance.

Phase 1: Initialization

The e-commerce platform is hosted and operated in a small shop and is not a critical AI system **(criticality level Low)**. No need for asset identification since there is only one AI asset, i.e. the chatbot.

The **trustworthiness dimensions** that are relevant and interested to the owners for this AI system (which is only 1 asset) are:

D1. Robustness – Ensuring the chatbot functions correctly under different conditions.

D2. Fairness & Bias – Avoiding discrimination in responses.

The assessment will occur at the **operation phase of this AI system**.

The organization contacted a co-creation workshop, used the TrustSense tool and it was concluded that the **AI team trustworthiness maturity (AIP) = Very High** since it demonstrates consistently high maturity regarding trustworthy.

From previous incidents that conducted criminal investigations the identified were youngsters with sophistication level **tA = Low**.

The risk appetite decided by the management of the shop: will treat 100% only risks scored Critical, Severe, High; will absorb all other risks.

Phase 2: Threat Assessment

Using OWASP AI we find that threats related to robustness, fairness and bias and the controls appropriate to mitigate these threats are listed:

Table 26: Sample of threats related to robustness, fairness, bias and the appropriate controls.

Threat	Description	Controls-Mitigation actions suggested by OWASP
Manipulation	Manipulating inputs to deceive AI models by crafting adversarial examples. Manipulating AI input prompts to bypass restrictions or extract sensitive data.	Adversarial training, input validation, model robustness techniques, anomaly detection. Input sanitization, prompt filtering, user access controls, sandboxed execution environments
Data Poisoning	Injecting malicious data into training datasets to corrupt AI learning.	**Data validation,** secure dataset curation, outlier detection, access control for training data.

		Efficient data management policy Use of tools
Model Evasion	Crafting inputs to bypass AI-based security mechanisms like spam filters.	Model hardening, behavioral analytics, continuous security assessments.
Model Extraction	Stealing AI models by making repeated queries to approximate its behaviour.	Rate limiting, **API security**, encrypted queries, watermarking models, model fingerprinting.
Model Inversion	Recovering private training data by exploiting model predictions.	Differential privacy, federated learning, **encryption techniques for sensitive data.**
Inherited bias from bias data	AI models inheriting biases from training data, leading to unfair decisions.	Bias detection tools, fairness-aware algorithms, diverse and representative training data Use of tools
Overloading	Overloading AI systems with excessive requests, causing slowdowns or crashes.	Rate limiting, request throttling, anomaly detection, server-side protections.
		T**ools** (see Table 13 in D.2,1 can be assessed and used to assess and mitigate the robustness and bias threats.

The controls that have already been implemented are the ones in bold.
Since the AI system is not critical we use the scale in Table 18:

Table 27: Measurement Scales.

Threats to the AI asset (chatbot)	**Threat Level**
Manipulation	Twice a year (Very High-**VH**)
Data Poisoning	Twice a year (Very High-**VH**)
Model Evasion	Once a year (High –(**H**))
Model Extraction	Once every 2 years –(Medium (**M**)
Model Inversion	Once every 5 years –(Low-**L**)
Inherited bias from bias data	Once every 10 years (Very Low-**VL**)

Overloading	Once every 5 years –(Low-**L**)

Phase 3 Impact Assessment

Table 28: Impact levels.

Threats	**Impact levels of the threats to the relevant trustworthiness dimensions**
Manipulation	The threat has **very serious consequences to dimensions D1 & D2 (I=VH)**
Data Poisoning	The threat has **serious consequences to D2 and some consequences to D1 (I=H)**
Model Evasion	The threat has **many consequences to D2 and some consequences to D1 (I=M)**
Model Extraction	The threat has **serious consequences to D2 and few consequences to D1 (I=H)**
Model Inversion	The threat has **serious consequences to D2 and some consequences to D1 (I=H)**
Inherited bias from bias data	The threat has **serious consequences to D2 and some consequences to D1 (I=H)**
Overloading	The threat has **very low consequences to D2 and serious to D1 (I=H)**

Phase 4: Vulnerability Assessment

Table 29: Vulnerability Level.

Threats	**Vulnerability Level of the AI asset to the threats**
Manipulation	None (0%) of the controls mentioned in first table have been implemented V=VH
Data Poisoning	Few (< 40-60%) of the controls have been implemented (V=M)
Model Evasion	None (0%) of the controls mentioned in first table have been implemented V=VH
Model Extraction	None (0%) of the controls mentioned in first table have been implemented V=VH**)**
Model Inversion	Few (< 40-60%) of the controls have been implemented (V=M)
Inherited bias from bias data	None (0%) of the controls mentioned in first table have been implemented V=VH**)**

Overloading	None (0%) of the controls mentioned in first table have been implemented V=VH**)**

Phase 5: Risk Assessment

Table 30: Risk estimation -1.

Threats	Threat Level	Vulnerability Level	Consequence Level	Risk Level
Manipulation	Very High	Medium	High	**High**
Data Poisoning	High	Very High	Medium	**High**
Model Evasion	Medium	Very High	High	**High**
Model Extraction	Low	Medium	High	**Minimal**
Model Inversion	Very Low	Very High	High	**Minimal**
Inherited bias from bias data	Low	Very High	High	**Medium**
Overloading	Very High	Very High	Very High	**Critical**
Manipulation	High	Very High	Medium	**Medium**
Data Poisoning	Medium	Very High	High	**High**
Model Evasion	Low	Medium	High	**Minimal**
Model Extraction	Very Low	Very High	High	**Medium**
Model Inversion	Low	Very High	High	**Medium**
Inherited bias from bias data	Very High	Very High	Very High	**Critical**
Overloading	Very High	Medium	High	**High**

Since the tAIP is Very High then Risk level will be reduced:

Table 31: Risk estimation -2.

Threats	Threat Level	Vulnerability Level	Consequence Level	Risk Level	Final Risk Level
Manipulation	Very High	Medium	High	High	**Medium**
Data Poisoning	High	Very High	Medium	High	**Medium**
Model Evasion	Medium	Very High	High	High	**Medium**
Model Extraction	Low	Medium	High	Minimal	**Minimal**

Model Inversion	Very Low	Very High	High	Minimal	**Minimal**
Inherited bias from bias data	Low	Very High	High	Medium	**Minimal**
Overloading	Very High	Very High	Very High	Critical	**High**
Manipulation	High	Very High	Medium	Medium	**Minimal**
Data Poisoning	Medium	Very High	High	High	**Medium**
Model Evasion	Low	Medium	High	Minimal	**Minimal**
Model Extraction	Very Low	Very High	High	Medium	**Minimal**
Model Inversion	Low	Very High	High	Medium	**Minimal**
Inherited bias from bias data	Very High	Very High	Very High	Critical	**High**
Overloading	Very High	Medium	High	High	**Medium**

Since the tAIP is Very High and the tA is low then the risk levels will be further reduced:

Table 32: Risk estimation -3.

Threats	**Threat Level**	**Vulnerability Level**	**Consequence Level**	**Risk Level**	**Final Risk Level**
Manipulation	Very High	Medium	High	High	**Medium**
Data Poisoning	High	Very High	Medium	High	**Medium**
Model Evasion	Medium	Very High	High	High	**Medium**
Model Extraction	Low	Medium	High	Minimal	**Minimal**
Model Inversion	Very Low	Very High	High	Minimal	**Minimal**
Inherited bias from bias data	Low	Very High	High	Medium	**Minimal**
Overloading	Very High	Very High	Very High	Critical	**High**
Manipulation	High	Very High	Medium	Medium	**Minimal**
Data Poisoning	Medium	Very High	High	High	**Medium**

Model Evasion	Low	Medium	High	Minimal	**Minimal**
Model Extraction	Very Low	Very High	High	Medium	**Minimal**
Model Inversion	Low	Very High	High	Medium	**Minimal**
Inherited bias from bias data	Very High	Very High	Very High	Critical	**High**
Overloading	Very High	Medium	High	High	**Medium**

Phase 6: Risk Management

Since the **risk appetite** decided by the management of the shop: will treat 100% only the risks scored critical, severe and high then we would propose to treat the:

Inherited bias from bias data	Very High	Very High	Very High	Critical	**High**

since this is the only one that is scored in this range.

Then we would propose to undertake the controls listed in the most updated lists of controls (e.g. OWASP). According to the initial table for this threat the proposed controls are:

"*Bias detection tools, fairness-aware algorithms, diverse and representative training data Use of tools*".