computers

Article

AI_TAF: A Human-Centric Trustworthiness Risk Assessment Framework for AI Systems

Eleni Seralidou [1,*], Kitty Kioskli [1], Theofanis Fotis [1,2] and Nineta Polemi [1,3]

[1] trustilio B.V., Vijzelstraat 68, 1017 HL Amsterdam, The Netherlands; kitty.kioskli@trustilio.com (K.K.); theo.fotis@trustilio.com (T.F.); nineta.polemi@trustilio.com (N.P.)
[2] School of Education, Sport & Health Sciences, University of Brighton, Brighton BN19PH, UK
[3] Department of Informatics, University of Piraeus, 185 34 Piraeus, Greece
[*] Correspondence: eleni.seralidou@trustilio.com

Abstract

This paper presents the AI Trustworthiness Assessment Framework (AI_TAF), a comprehensive methodology for evaluating and mitigating trustworthiness risks across all stages of an AI system's lifecycle. The framework accounts for the criticality of the system based on its intended application, the maturity level of the AI teams responsible for ensuring trust, and the organisation's risk tolerance regarding trustworthiness. By integrating both technical safeguards and sociopsychological considerations, AI_TAF adopts a human-centric approach to risk management, supporting the development of trustworthy AI systems across diverse organisational contexts and at varying levels of human–AI maturity. Crucially, the framework underscores that achieving trust in AI requires a rigorous assessment and advancement of the trustworthiness maturity of the human actors involved in the AI lifecycle. Only through this human-centric enhancement can AI teams be adequately prepared to provide effective oversight of AI systems.

Keywords: Artificial Intelligence; trustworthiness; human-centric; framework

Academic Editors: Amit Kumar Mishra and Deepak Puthal

Received: 11 May 2025
Revised: 18 June 2025
Accepted: 19 June 2025
Published: 22 June 2025

Citation: Seralidou, E.; Kioskli, K.; Fotis, T.; Polemi, N. AI_TAF: A Human-Centric Trustworthiness Risk Assessment Framework for AI Systems. *Computers* **2025**, *14*, 243. https://doi.org/10.3390/computers14070243

1. Introduction

Artificial Intelligence (AI) has made profound advancements across critical sectors of society, including healthcare, law enforcement, transportation, and public governance. The integration of AI into decision-making processes in these fields has transformed operations from predictive policing and autonomous vehicles to AI-driven diagnostics and resource management in public administration [1]. As AI becomes increasingly embedded in high-stakes decisions, the need for trustworthy AI systems is growing. Trustworthiness in AI ensures not only operational efficiency and robustness but also safeguards ethics and privacy, prevents biases, and ensures accountability, accuracy, and transparency. The National Institute of Standards and Technology (NIST) introduced the AI Risk Management Framework (AI RMF), which provides a comprehensive approach to managing AI deployment risks while enhancing transparency and system integrity [2].

The European Union (EU) legislation, the AI Act [3], categorises AI systems according to their trustworthiness risk levels and imposes strict requirements (e.g., human oversight) on high-risk applications [4].

While these regulatory frameworks provide essential technical and procedural foundations, ensuring trust in AI systems depends equally on human involvement. Building organisational capacity through the establishment of skilled multidisciplinary teams is crucial for overseeing the appropriate application of AI.

This paper introduces an AI trustworthiness risk assessment framework (AI_TAF), which provides a structured, lifecycle-oriented methodology tailored to be both human-centric and adaptable across different sectors. The framework underscores the need to evaluate AI-specific vulnerabilities—such as bias, opacity, and robustness deficiencies—across all key assets of an AI system, including training models, algorithms, and datasets, throughout each phase of its lifecycle (design, development, integration, and deployment stages).

This paper is structured as follows: In Section 1, we provide the introduction. Section 2 reviews the background and related works. Section 3 focuses on the human element of AI trustworthiness, including team roles, attributes, and maturity assessments. Section 4 presents the AI_TAF risk management framework, detailing its six phases, from initialisation to risk management, along with an example of its application. Section 5 outlines directions for further research. Finally, Section 6 concludes the paper.

2. Background and Related Work

The concept of trustworthy AI includes dimensions like transparency, fairness, robustness, and accountability [5]. These dimensions echo those established by the AI HLEG (High-Level Expert Group) adopted in the AI Act [6] and refer to the following guiding principles in various Articles of the AI Act: data and data governance (Article 10), technical documentation (Article 11), record-keeping (Article 12), transparency (Article 13), human oversight (Article 14), accuracy, robustness, and cybersecurity (Article 15).

When considered together, these elements form a holistic foundation that enhances the reliability, efficiency, and societal trustworthiness of AI systems (Figure 1), with human oversight being a horizontal obligation.

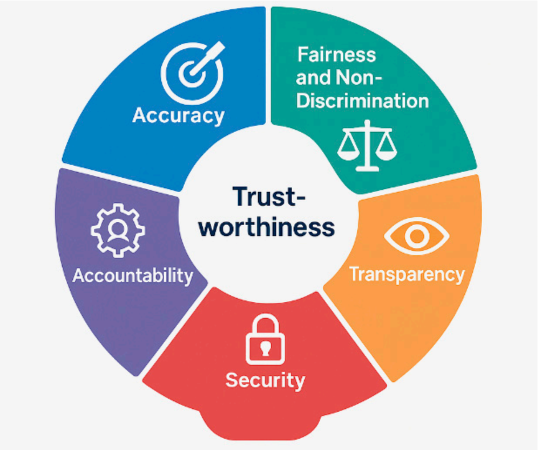


Figure 1. Trustworthiness dimensions.

A core requirement of the AI Act is the implementation of a **trustworthiness risk management system**, as outlined in Article 9. This entails a continuous, lifecycle-wide process aimed at identifying threats, vulnerabilities, and potential impacts across key dimensions of trustworthiness. The objective is to ensure that AI systems are designed and developed in accordance with the essential requirements for reliability, safety, and ethical integrity.

Establishing trustworthiness requirements is central to emerging certification schemes for AI systems [7] and will lead to AI certification, which is another requirement of the AI Act in Article 43.

As highlighted by Jobin et al. [8], more than 80 ethical guidelines on AI have been proposed globally, underscoring the widespread consensus on the need for human-centred AI development. These guidelines emphasise the importance of developing AI systems that align with human values and prioritise societal well-being [9], marking a global shift toward ensuring that AI serves humanity's best interests.

Recent studies have emphasised that technical measures alone are insufficient to ensure secure ICT systems or trustworthy AI [10,11]. As Hagendorff [12] and Mökander and Floridi [13] note, building trust in AI requires more than just improving the algorithms. The key additional elements include the following:

- Contextual Understanding: AI systems must be developed with an understanding of the specific social, cultural, and environmental contexts in which they will be deployed. This contextual awareness helps avoid unintended consequences and ensures that AI is used appropriately in various scenarios.
- Multi-Stakeholder Engagement: The development of AI must involve a range of stakeholders, from engineers and policymakers to the communities impacted. Engaging diverse groups ensures that AI systems reflect the interests and needs of all stakeholders.
- Ongoing Assessment: Continuous monitoring and evaluation of AI systems are necessary to address emerging risks and ensure that the systems remain aligned with ethical and legal standards. Regular audits and updates of AI models help to maintain accountability and trust.

Several standards and initiatives have been proposed to assess and mitigate the risks associated with AI. These include:

- NIST AI RMF [2]: The National Institute of Standards and Technology (NIST) developed the AI Risk Management Framework (RMF) to guide organisations in managing the risks associated with AI systems. This framework emphasises transparency, accountability, fairness, and robustness.
- ISO/IEC 23894 AI Risk Standard [14]: This international standard focuses on providing guidelines for managing AI risks, with particular emphasis on the safety and security of AI systems and ensuring that they function as intended without causing harm.
- IEEE Trustworthiness Guidelines (IEEE, 2024) [7]: These guidelines aim to ensure that AI technologies are trustworthy by outlining principles such as fairness, transparency, and accountability. They offer ethical guidance for the responsible development of AI technologies.
- ISO27090 [15], ISO27091 [16], ISO/IEC 5338 [17] evaluation of AI threats and definition of AI lifecycle are included in these standards.
- ETSI TC SAI Activity Report [18]: four reports collectively address the explicability and transparency of AI processing and provide an AI computing platform security framework; threats posed by so-called 'deepfakes' and strategies to minimise them are included.
- ISO/IEC42001 [19] specifies the requirements for establishing, implementing, maintaining, and continually improving an AI Management System within organisations.
- The ENISA's FAICP framework [20]: a framework for good AI cybersecurity practices necessary for securing ICT infrastructures and hosted AI, considering the AI life cycle, which goes beyond ML (from system concept to decommissioning) and all elements of the AI supply chain, associated actors, processes, and technologies.
- The OWASP AI Exchange [21] is a comprehensive core framework of cybersecurity and privacy threats, controls, and related best practices for all AI, which is actively aligned with international standards and fed into them.

However, many of these frameworks, while comprehensive, lack practical tools to assess team readiness and evaluate internal organisational processes for implementingthe guidelines and do not consider the maturity of the AI teams to oversee the AI systems (e.g estimating, and mitigating risks) [22,23]. This gap implies that organisations may struggle to apply these frameworks effectively in practice. The absence of structured mechanisms to assess the trustworthiness maturity of human actors overseeing AI development means that risks linked to inadequate oversight, misalignment with ethical principles, or poor cross-functional collaboration may go undetected. Unlike threats to data or algorithms, these human-centric vulnerabilities are harder to monitor yet often lead to substantial failures of trust. By "practical tools", we refer to applied, operational instruments—such as structured maturity evaluation schemes, behavioural scoring models, and phase-specific team capability assessments— that can be systematically applied across lifecycle stages. These are not merely checklists but mechanisms embedded in the risk assessment itself. Contemporary literature acknowledges that technical safeguards alone are insufficient unless accompanied by institutional and human readiness to enforce and interpret them effectively [12,13].

There are various open-source tools that assess threats that impact various AI trustworthiness dimensions, e.g., Table 1.

Table 1. Open-source AI assessment tools.

Open-Source AI Assessment Tools	Trustworthiness Dimension	Features
AIF360 (IBM) [24]	Fairness & Bias	Bias detection metrics, bias mitigation algorithms
AIX360 (IBM) [25]	Explainability	LIME, SHAP, prototype-based explanations, global/local interpretability
Adversarial Robustness Toolbox (ART) [26]	Robustness & Security	Simulates adversarial attacks and defenses across ML frameworks
Fairlearn [27]	Fairness & Bias	Fairness metrics and reduction-based bias mitigation algorithms
What-If Tool [28]	Explainability & Fairness	Visual model analysis, counterfactuals, slicing
DeepChecks [29]	Model Validation & Testing	Bias detection, data drift, leakage, performance issues
Model Card Toolkit [30]	Transparency & Documentation	Generates model documentation (intended use, metrics, limitations)

Organisations can use these tools as technical controls against threats like lack of fairness, transparency, explainability, and robustness of security.

3. Human Element in the AI Trustworthiness

It is well acknowledged that human involvement remains one of the most critical yet complex components influencing the cybersecurity of ICT systems [10,11,31].

In the abovementioned initiatives, while much attention has been given to technical robustness, data quality, and algorithmic transparency, human participants in the AI lifecycle are not considered in the trustworthiness of the system. The AI_TAF fills this gap by prioritising the human dimension in trustworthiness evaluations.

3.1. Teams and Participants in the AI Lifecycle

Human involvement remains one of the most critical and complex components influencing the trustworthiness of AI systems. The human participants—designers, developers, testers, domain experts, operators, users, and decision-makers (Table 2)—in all phases of the AI lifecycle (design, development, and deployment phases) play a pivotal role (Figure 2).

Table 2. AI participants and their roles in AI teams.

AI Participants	Roles and Responsibilities
Data Scientists/ AI Engineers	Develop/train AI models; data processing, model selection, training/evaluation/optimisation.
Domain Experts	Provide specialised 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, and 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	Utilise AI systems in business sectoral applications; provide feedback on system performance, report issues, and contribute to systems 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.

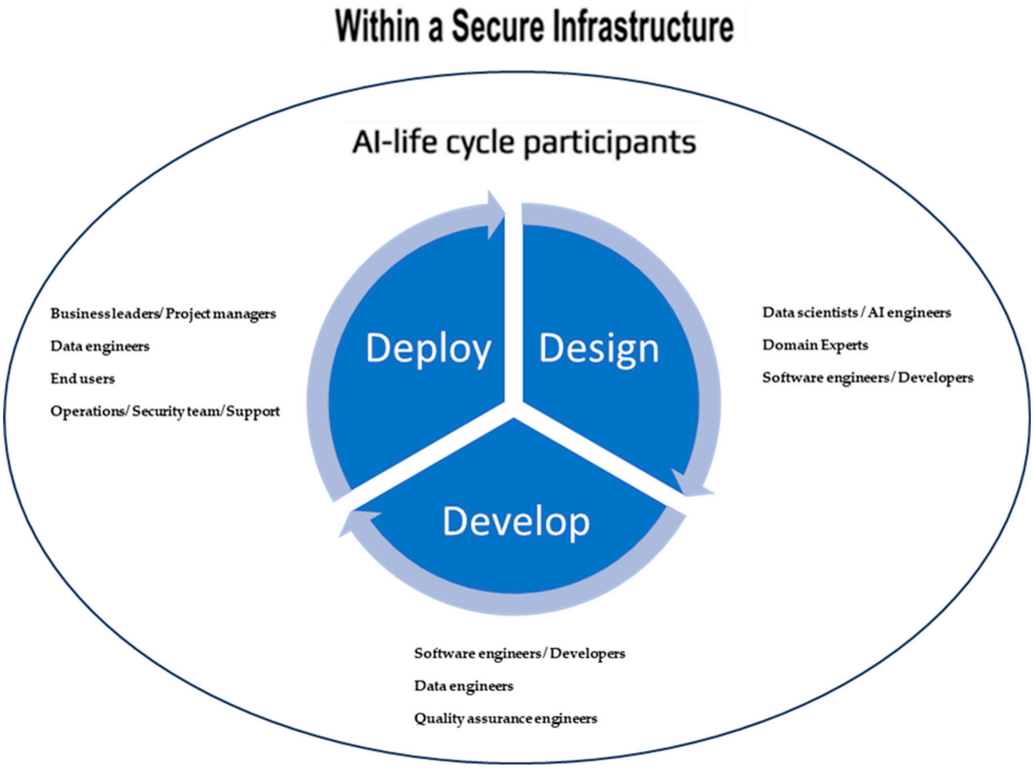

Within a Secure Infrastructure
AI-life cycle participants
Deploy
Design
Develop
Business leaders/ Project managers
Data engineers
End users
Operations/ Security team/Support
Data scientists / AI engineers
Domain Experts
Software engineers / Developers
Software engineers / Developers
Data engineers
Quality assurance engineers

Figure 2. AI participants.

AI teams consist of participants with different roles and responsibilities.

Fostering trust in an AI system, even when it operates within a secure infrastructure, extends beyond technical protection. It also requires the establishment of capable and competent human teams to oversee AI systems. Human oversight—a horizontal requirement under the AI Act—entails the ability to intervene, halt, or override AI operations when required. This oversight must be conducted by AI professionals who can:

- Prevent or mitigate risks;
- Promptly interrupt or stop system operations;
- Override AI decisions that may result in harmful or unjust outcomes;
- Identify and correct errors;
- Understand the AI system's functionality and limitations;
- Clearly explain the rationale behind AI-generated decisions;
- Respond to and report incidents.

The maturity and readiness of the AI teams participating in the AI lifecycle are integral to ensuring the overall trustworthiness of AI systems.

Trustworthy AI systems are fundamentally shaped by the competence, value, and situational awareness of their users. This includes their ethical maturity, personality traits, cognitive capabilities, and resilience under pressure. Attributes such as vigilance, adaptability, critical thinking, and interdisciplinary collaboration are essential when humans must design, test, deploy, interpret, or respond to AI outputs, especially in high-stakes or fast-paced environments [32].

Human readiness is not a static trait but a dynamic condition that must be continually assessed and nurtured through various types of participation in all phases of the AI lifecycle. Factors such as ethical awareness, technical proficiency, and exposure to AI literacy programmes significantly affect how individuals engage with AI. Cross-functional collaboration—bringing together technologists, domain experts, ethicists, and users—supports more holistic decision-making and mitigates siloed thinking that can compromise system integrity [33].

Moreover, the organisational context plays a vital role. Organisational culture, governance structures, and resource availability influence whether individuals are empowered to raise concerns, apply ethical principles, and act responsibly in the face of uncertainty or failure. The presence of necessary resources shared accountability mechanisms, and clear escalation pathways further support human contributions to trustworthy AI [34].

Equally important is the impact of human cognitive limitations. Trust can be undermined by biases (e.g., automation or confirmation bias), decision fatigue, over-reliance on AI recommendations, and misinterpretation of system capabilities. These risks are particularly acute when humans are expected to serve as the "last line of defense" against erroneous AI behaviour without the necessary support, training, or situational awareness [35].

Furthermore, adversaries can exploit human vulnerabilities as part of broader socio-technical threats. Malicious actors, from hacktivists to state-sponsored cyber agents, often use deception, misinformation, phishing, or social engineering to manipulate human operators or corrupt system behaviour. Understanding attacker profiles and motivations can provide a more realistic view of system vulnerabilities and their severity, thereby enhancing organisational preparedness [9]. Moreover, a human-centric approach is essential for promoting cybersecurity hygiene, particularly in critical sectors such as healthcare, where trust and safety are paramount [10,11].

In summary, the human element in AI trustworthiness is multifaceted and involves personal, organisational, and socio-technical layers. To address this, the AI_TAF framework promotes the ongoing evaluation of human maturity and proposes social measures, including proactive education in AI ethics and critical thinking, behavioural change interventions, co-creation workshops, and the development of environments that support ethical decision-making, collaboration, and psychological resilience in the use of AI. The goal of trustworthy AI can only be fully realised by acknowledging and addressing human vulnerabilities.

The AI_TAF emphasises that trust in AI cannot be fully realised without a thorough evaluation and strengthening of the human dimension throughout the AI lifecycle. Only by doing so can individuals be truly equipped to exercise effective oversight of AI systems (as imposed by the AI Act, Article 14).

3.2. AI Teams' Attributes and Trustworthiness Maturity

For AI initiatives to succeed within organisations, both the composition of the AI teams, broader organisational context, and intended use of the AI system play crucial roles. A mature organisation that aims to adopt AI must foster specific team characteristics and structural capacities that support not only the design and deployment of AI systems but also their continuous improvement and ethical use [36].

AI team attributes extend beyond technical expertise. Successful AI teams typically blend diverse expertise, including data science, domain knowledge, software engineering, user experience, and ethical governance. Multidisciplinary collaboration enables better alignment with organisational goals and user needs. Furthermore, AI teams must exhibit adaptability, continuous learning, and the capacity for critical reflection, especially when dealing with complex data and emerging technologies [37].

Effective teams also embrace agile and iterative approaches. This approach helps them remain responsive to changes in data availability, user feedback, and the evolving regulatory landscapes. Strong internal and external communication skills are essential, particularly when engaging stakeholders or translating technical developments into strategic values.

Regarding organisational maturity, the readiness to integrate AI depends heavily on the presence of foundational enablers. These include established data governance policies, a robust IT infrastructure, clear leadership support, and a culture of innovation and trust [38]. Mature organisations cultivate environments where experimentation is encouraged, failures are viewed as learning opportunities, and cross-departmental collaboration becomes routine.

Moreover, mature organisations tend to embed AI into their strategic vision. They support AI teams not only with resources but also with ethical frameworks, training programmes, and decision-making autonomy. Maturity is also reflected in the ability to scale AI efforts, moving from pilot projects to enterprise-wide implementations while managing risks and ensuring accountability [39].

In summary, high-functioning AI teams and a mature organisational context are deeply intertwined (Figure 3). One cannot thrive without the other.

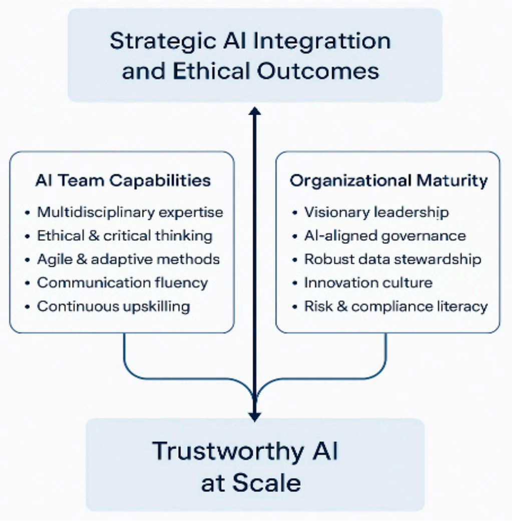

Strategic AI Integrattion and Ethical Outcomes
AI Team Capabilities
• Multidisciplinary expertise
• Ethical & critical thinking
• Agile & adaptive methods
• Communication fluency
• Continuous upskilling
Organizational Maturity
• Visionary leadership
• AI-aligned governance
• Robust data stewardship
• Innovation culture
• Risk & compliance literacy
Trustworthy AI at Scale

Figure 3. Dual engine model for sustainable AI integration.

Organisations that understand and nurture this symbiotic relationship are better positioned to leverage AI for meaningful and sustainable transformations.

3.3. Trustworthy Maturity Evaluation of AI Teams (tAIP—Trustworthy AI Participants)

The Trustworthy AI Maturity for Teams (tAIP—trustworthy AI Participants) measures the effectiveness of practices in identifying trustworthiness threats, evaluating and mitigating AI trustworthiness risks, and focusing on the roles and behaviours of AI-enabled teams. To account for the varying levels of technical proficiency among different participants in the AI lifecycle (Figure 2), two related questionnaires are proposed. These questionnaires are structurally similar but differ primarily in the Technical Proficiency section, which is tailored separately for Technical AI participants (e.g., developers and engineers) and Non-Technical participants (e.g., domain users).

Assessing AI participants' trustworthiness is more relevant in sectors with diverse, unvetted AI users, while in contexts with trained personnel, organisational assessments may already cover these aspects. In such cases, assessing AI maturity at the team or organisational level may be more appropriate.

The process for measuring the Trustworthy AI Maturity for Teams involves evaluating the AI team's maturity across several key dimensions, such as proactivity, responsibility, ethics, innovation, resilience, collaboration, technical proficiency, and compliance. The evaluation was conducted by an AI risk manager who used a Likert scale to systematically evaluate the team's agreement with a set of structured statements that represented these dimensions. Each dimension was assigned a weight based on its relevance to the organisation. The team's responses are scored and then averaged to produce dimension scores. These scores are weighted and combined to generate an overall trustworthiness score for the team. The score is then categorised into levels (e.g., Very High, High, Moderate, etc.), which provides a clear indication of the team's AI maturity and trustworthiness. This process helps organisations identify areas for improvement and ensures that AI systems are managed with an appropriate level of trust.

The organisation can use the overall score of the maturity level of their AI teams (tAI) to follow the mitigation suggestions outlined in the following table (Table 3).

Table 3. Maturity levels of AI teams (tAIP) and mitigation actions.

Levels (tAIP)	Scoring	Analysis of Findings	Suggested Mitigations for Risk Reduction
Very High (VH)	5	The team generally meets trustworthy AI requirements, in all areas	To sustain this level, regularly conduct team training, celebrate collective achievements, and foster a culture of continuous improvement.
High (H)	4	The team generally meets trustworthy AI requirements, with a few areas that need improvement.	Improve organisational training, encourage collaboration between teams, and refine adherence to ethical standards and company policies to enhance performance.
Substantial (S)	3	The team partially meets trustworthy AI requirements, highlighting areas that require attention.	Introduce structured training programmes, strengthen team collaboration, and promote mentorship within the organisation to address these gaps. Adopt an access and logging policy.
Medium (M)	2	There are significant gaps in the team's trustworthy AI maturity.	Organise intensive workshops, focus on ethical compliance, and implement policies to improve trustworthiness practices across teams. Adopt a strict access, logging policy and a least privilege approach.
Low (L)	1	The team faces substantial challenges in achieving trustworthy AI maturity.	Commit to comprehensive retraining, track team progress through assessments, and implement supervised practices to restore essential trustworthiness traits. Do not allow access to model; special permission is required.

AI_TAF uses the tAIP scores to adjust the default trustworthiness risk estimations, as discussed in Section 4. For each AI asset under assessment, the team's "owner of the asset" is identified, meaning the team with AI participants (see Figure 2) that interacts (designs/develops/integrates/uses/operates) with the asset. It may be only one team that owns the asset (s), or it may be one AI participant in the team. The tAIP is estimated for each team owner.

Additionally, the dimensions used in the tAIP model draw on interdisciplinary perspectives from AI governance, organisational behaviour, and responsible innovation. Constructs such as ethical awareness, resilience, and collaborative maturity have been identified in prior studies as foundational to trustworthy human−AI interactions [5,8]. While the scoring system is currently heuristic, it was designed to align with the operational realities faced by AI teams across sectors. Future versions of AI_TAF will aim to empirically validate these dimensions using behavioural and organisational metrics to enhance their consistency and objectivity.

3.4. Sophistication of Potential AI Adversaries (tA)

In the AI_TAF framework, assessing the maturity of potential adversaries is optional and may be considered if organisations have advanced cybersecurity intelligence capabilities, such as historical data on adversaries, such as from past cybercriminal investigations or from the MITRE ATT&CK database [40].

This includes tracking digital footprints, analysing behaviours, and understanding the motivations of these individuals. A profiling scale for potential AI adversaries has been developed, like the AI teams' trustworthiness maturity estimation. This scale helps organisations identify internal adversaries and assess their potential to carry out sophisticated attacks [9,41].

Each question in the adversary profiling scale is rated on a 5-point Likert scale, and the responses are averaged to produce a final score.

The scores for each section (e.g., Technical Traits) are summed and converted into a percentage. This percentage is then compared to predefined ranges to determine the adversary's profile, such as "Experienced" or "Novice," based on the total score. This process helps identify the sophistication of potential adversaries and guides defensive strategies.

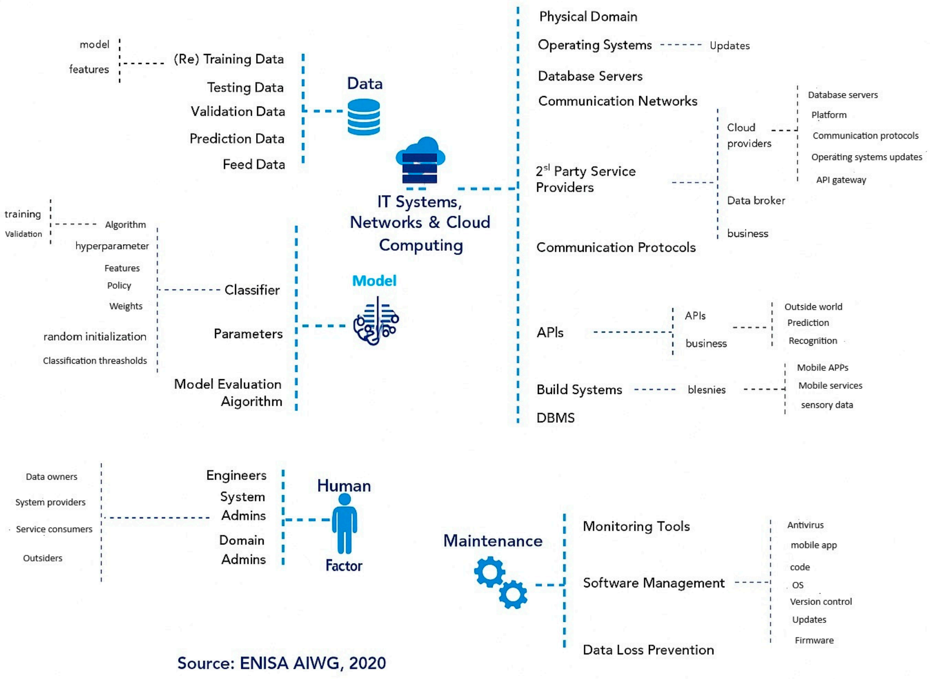
4. AI_TAF Introduction and Objectives

The proposed AI trustworthiness risk assessment framework (AI_TAF) entails a continuous, repeated, lifecycle-wide process aimed at identifying threats, vulnerabilities, and potential impacts in each stage of the AI lifecycle across the key dimensions of trustworthiness of all AI system assets (components) [42] involved in the assessment stage (Figure 4).

It combines the principles of ISO27005 [43] risk management with maturity assessment to guide informed decision-making for each stage of the AI lifecycle for each AI system's asset under assessment.

AI_TAF adopts a unified approach to trustworthiness where the "AI System Trustworthiness" interlinks with the "AI team Trust Maturity" (Figure 5), ensuring that AI systems—especially high-risk ones—are designed and used in ways that allow humans to understand, monitor, and intervene; the humans are capable of responding to this need.

Unlike existing frameworks that often treat trustworthiness as a property of the system or data, AI_TAF reconceptualises trust as a joint property of the AI system and its human stewards. By systematically quantifying the Trust Maturity of AI teams (via tAIP) and incorporating these scores into the risk model, AI_TAF not only acknowledges but also operationalises the human dimension of trust. This goes beyond "human oversight" checklists by embedding team readiness, ethics, and cross-functional collaboration into the risk equation, making the framework genuinely human-centric in its design and application.



Source: ENISA AIWG, 2020

Figure 4. AI system assets.

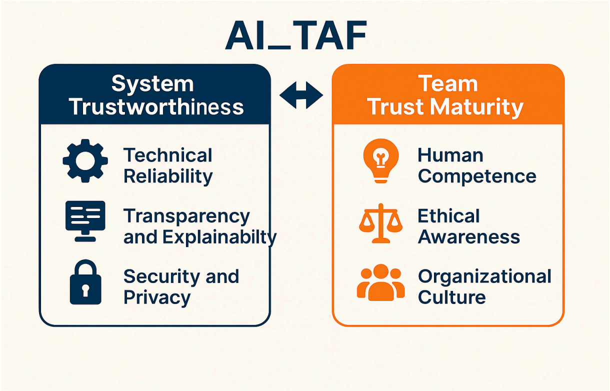


Figure 5. AI_TAF links "System Trustworthiness" and "Team Trust Maturity".

The framework aims to classify a broad range of threats across all key dimensions of trustworthiness, drawing on the established initiatives mentioned in Section 2. It also identifies the existing challenges in fully mapping the AI threat landscape and highlights areas requiring further research to enhance trust and resilience.

AI_TAF approaches trustworthiness as a dynamic attribute, not something achieved at launch but maintained through continuous assessment. AI_TAF can be applied to each stage of the AI lifecycle. Each assessment incorporates indicators for these attributes (stage of the lifecycle, AI assets involved, AI team "owners" of each AI asset, and trustworthiness dimensions relevant to the stage), linking them to specific threats and mitigations. Hence, the AI_TAF will be applied iteratively to evaluate threats and estimate risks for all assets of the AI system throughout its entire lifecycle, considering sector-specific characteristics, the intended use of the AI system, the trustworthiness of the teams involved, and (optionally) the sophistication level of the potential AI adversaries.

4.1. Guided Principles and Assumptions

The framework is guided by the following foundational principles:

- Inclusivity: Leverages established risk assessment methodologies to build on well-tested practices (e.g., it is based on the NIST AI-RFM).
- Human-Centric Focus: Considers the reliability and roles of human actors in AI systems.
- Lifecycle Coverage: Applies to every stage, from design to deployment and operation.
- Tool Neutrality: Operates independently of specific technologies or platforms.
- Sector Adaptability: Accommodates diverse industry contexts.
- Standards Compliance and Global Alliance: Aligns with international norms like ISO 27005 [43], ISO 42001 [19], ENISA FAICP [20], and NIST AI RMF [2] (Section 2).
- Cross-Domain Usability: Supports a wide array of stakeholders, regardless of their field.

Expanding on the ENISA FAICP [20] model, AI_TAF positions AI systems as part of broader ICT ecosystems and introduces a three-layered structure for assessing trustworthiness:

4.1.1. Layer I: Foundational Assumptions

This layer outlines the essential conditions presumed during the assessments.

- The digital infrastructure that hosts the AI system under assessment is secure, i.e., it adheres to cybersecurity best practices, e.g., ISO 27001 [44] certified.
- The data sources are certified (the integrity of the training data is ensured), minimising threats like data poisoning.
- Each AI asset under assessment has an identified "owner" (team of AI participants) who is not necessarily distinct and is responsible for AI asset oversight.
- A designated individual (e.g., a risk assessor) leads the trustworthiness evaluation, supported by input from others. Under the consensus of the AI asset owners, the risk assessor contacted workshops to anonymously evaluate and provide the tAIP scores of each AI asset owner. Optionally, the risk accessor provides the sophistication score (tA) of the potential adversaries.
- The AI system is either securely isolated or connected to an equally protected environment.

4.1.2. Layer II: General AI Trustworthiness Assessment

This layer focuses on evaluating the distinctive risks of AI assets at each stage of the lifecycle. It provides a trustworthiness assessment based on ISO27005 [41] without considering the environmental specificities.

4.1.3. Layer III: Sector-Specific Environmental Assessment

This final layer adjusts the assessment considering the specificities of the operational environment, which include the following:

- Criticality of the sector that the AI system is being used
- Intended use of the system
- AI team composition and responsibilities
- Maturity level of the AI team towards trustworthiness
- Potential Adversaries and level of sophistication
- Business objectives and trustworthiness risk appetite (the amount and type of risk that an organisation is willing to accept in pursuit of its objectives).

Based on the environment, the risk evaluation will be adjusted, and the proposed controls will be tailored to the specific characteristics of the environment.

While the AI_TAF framework draws structural inspiration from established standards such as ISO 27005 and the NIST AI Risk Management Framework, its contribution lies in the integration of human-centric elements into the core methodology. Specifically, AI_TAF introduces the Trustworthy AI Maturity for Teams (tAIP) model, which dynamically influences risk prioritisation across the AI lifecycle. In contrast to existing frameworks, which emphasise system properties and organisational compliance, the AI_TAF assigns trust risk not only to AI assets but also to their human owners. This dual-layered approach allows for more granular, phase-specific evaluations and addresses gaps in operational readiness and team accountability that are not explicitly covered in the current standards.

4.2. *AI_TAF: A Six-Phase Approach to AI Risk Management*

The AI_TAF presents a well-organised six-phase framework (based on ISO27005) for managing AI-related risks (Figure 6); it is iterative, supporting the assessment of AI systems for each stage of their lifecycle.

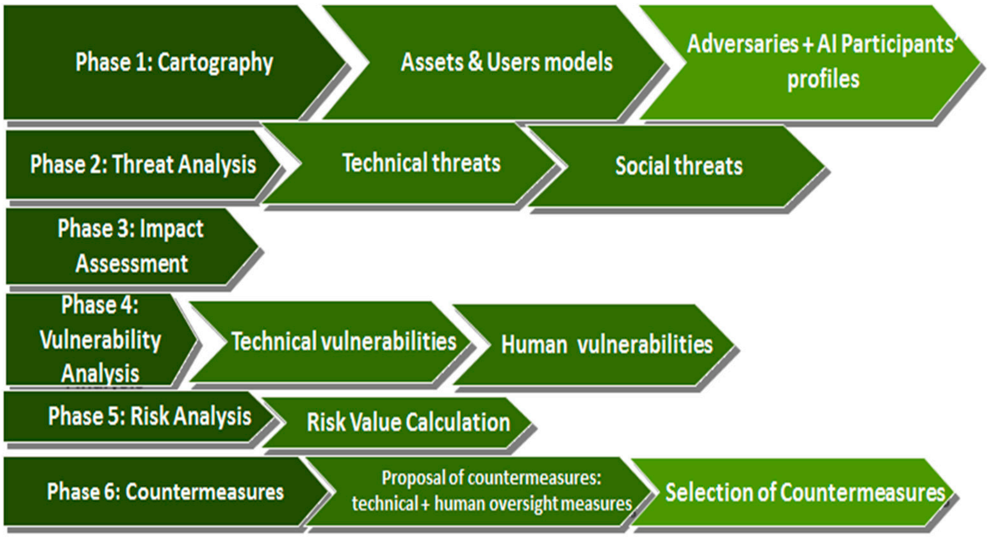

Phase 1: Cartography
Assets & Users models
Adversaries + AI Participants' profiles
Phase 2: Threat Analysis
Technical threats
Social threats
Phase 3: Impact Assessment
Phase 4: Vulnerability Analysis
Technical vulnerabilities
Human vulnerabilities
Phase 5: Risk Analysis
Risk Value Calculation
Phase 6: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

Figure 6. AI_TAF phases.

4.2.1. Phase 1: Cartography-Initialisation

It begins with clearly establishing the scope of the assessment and setting the initialisation attributes. This involves identifying the AI system, the stage of the AI lifecycle, clarifying its intended purpose, mapping out the AI team involved in this stage of the lifecycle under assessment, and cataloguing the AI system's assets to be assessed and the controls that have already been implemented to ensure its trustworthiness.

The criticality of an AI system (as outlined in Table 4), proposed in this paper, is determined by several key factors that help determine the overall importance and regulatory sensitivity of the system, guiding the depth and focus of the trustworthiness and risk assessment.

The criticality level of an AI system imposes the same criticality level on all its assets under assessment.

To support this process, an *asset model* is created, offering a structured representation of the assets of the AI system in this phase of the lifecycle. This model (e.g. Figure 7) highlights the relationships and dependencies among assets, such as training models, algorithms, datasets, workflows, and human actors.

Table 4. AI system criticality level.

Criticality Level	Criteria for Determining Criticality of an AI System
Very High (VH)	- The system is classified as high-risk according to Article 6 of the AI Act, including use cases listed in Annex III. - It poses an extreme threat to human health, safety, or fundamental rights—this may apply even if the system does not have a direct influence on decision-making outcomes.
High (H)	- The system presents a significant threat to health, safety, or individual rights, even without materially influencing decisions. - It is deployed by an Operator of Essential Services (OES) and is used to deliver services listed in Annex I of the NIS2 Directive (essential services).
Substantial (S)	- The system carries a notable risk of harm, including influencing decision-making that may affect health, safety, or basic rights. - It is utilised by an OES to support the delivery of an important service, as outlined in Annex II of the NIS2 Directive.
Medium (M)	- The AI system entails a moderate level of risk to individuals' rights or well-being and may shape decision-making outcomes. - It is involved in providing an important service (as defined in Annex II of the NIS2 Directive), though it may not be operated by an OES.
Low (L)	- The system poses little to no risk to human health, safety, or fundamental rights, including when it influences decision-making. - Typically, such systems require minimal regulatory intervention and lighter trustworthiness assurance.

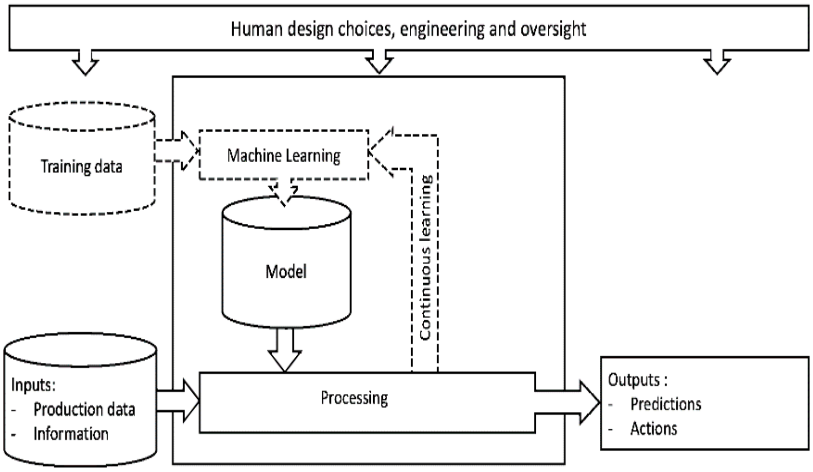


Figure 7. AI asset model.

By visualising these interconnections, the asset model not only enhances understanding of the system's structure but also helps pinpoint which elements are most critical to evaluate. Ultimately, this facilitates a more focused and effective prioritisation of trustworthiness concerns across the AI lifecycle.

All implemented controls (technical, procedural, and business) to protect the AI assets under assessment against AI threats must be listed.

A user model will be created, outlining the AI teams involved in this phase of the lifecycle, which are the owners of the AI assets under assessment.

The assessor will contact workshops to evaluate the maturity level of all owners towards trustworthiness and provide tAIP scores. Optionally (upon threat intelligence capabilities), the sophistication scores of potential adversaries (tA) are also provided.

Phase 1 Output:

All attributes for the initialisation of AI_TAF will be provided, namely, the criticality of the AI system, the stage of lifecycle considered in this assessment, the AI asset model; implemented controls; AI user model; maturity scores of the teams (tAIP); and optionally, the sophistication scores of the adversaries (tA).

4.2.2. Phase 2: Threat Assessment

Threat assessment involves a thorough evaluation of potential threats to all AI assets in this phase of the AI lifecycle that could compromise system trustworthiness. This includes technical threats such as malware and data poisoning, as well as social threats like phishing and social engineering, which may vary depending on the AI maturity level of the organisation's teams. The assessment identifies threats related to data quality, AI model performance, and operational deployment issues. The OWASP AI exchange [21] and ENISA can be used to identify such threats.

During this phase, we also estimate the likelihood of each AI threat for each AI asset under assessment occurring (Table 5) (as recorded by administrators or through logs). Factors influencing threat occurrence include:

- Historical Data: Past occurrences of similar threats help forecast future risks;
- Environmental Factors: Conditions in the sector or location where the AI system operates, such as natural disasters or political/economic stability;
- Stability and Trends: Geopolitical events (e.g., economic crises, wars, and pandemics) and technological advancements (e.g., AI-driven attack systems) may signal an increase in threats.

Table 5. Threat level scale.

Threat Level	Occurrence Rate
(VH) = 5	Twice a year
(H) = 4	Once a year
(S) = 3	Once every 2 years
(M) = 2	Once every 5 years
(L) = 1	Once every 10 years

The proposed scale in Table 5 can be adjusted based on the criticality of the AI system and the organisation's "risk appetite". For example, for AI systems with a Very High criticality level, the threat level is categorised as Very High if it occurs twice within the last 10 years.

Phase 2 Output:

All technical and social threats have been identified, and the likelihood of each threat occurring to the AI components of the system under evaluation has been assessed.

4.2.3. Phase 3: Impact Assessment

The goal is to assess the potential impact of each threat on the different dimensions of trustworthiness relevant to the asset and phase in the lifecycle under assessment. This phase involves a thorough evaluation of how identified threats could affect these trustworthiness dimensions across. By referencing repositories like OWASP and ENISA, we can pinpoint the overlapping consequences of threats. For instance, threats like data loss and model poisoning can negatively impact multiple dimensions of trustworthiness, such as accuracy, fairness, and cybersecurity. Another dimension explicitly considered in AI_TAF is explainability. Threats that affect transparency—such as model opacity, biased training logic, or black-box decision-making—can compromise the ability of human actors to interpret system behaviour. This is particularly relevant in high-stakes environments where decisions must be justified or, at times, overridden. As part of the impact evaluation, explainability is assessed as a distinct dimension whenever it is applicable to the AI asset under evaluation.

The table below (Table 6) will be adjusted to reflect the specific dimensions of trustworthiness pertinent to the system being assessed. The impact on each dimension may differ, and an average will be calculated. Additionally, the impact of each threat can be further evaluated based on the environment in which the AI system operates. In this case, business, technological, legal, financial, and other consequences need to be evaluated that organisations may encounter if the affected trustworthiness dimensions are compromised.

Table 6. Impact of each threat on trustworthiness dimensions.

Consequence (Impact) Level	Means
(VH) = 5	The threat has significant and severe impacts on all trustworthiness dimensions, affecting them in multiple ways.
(H) = 4	The threat has substantial impacts on most trustworthiness dimensions.
(S) = 3	The threat has moderate impacts on many trustworthiness dimensions.
(M) = 2	The threat has minimal impacts on some trustworthiness dimensions.
(L) = 1	The threat has negligible impacts on any of the trustworthiness dimensions.

Output of Phase 3:

The consequences of each threat in relation to the trustworthiness dimensions will be evaluated. These assessments will include numerical scores, and, where necessary, qualitative reports will be provided to offer detailed insights into the assessment findings, such as the nature of the threats. Business impact assessment reports reveal the various consequences (e.g., legal, business, financial, technological, societal, and reputation) that the affected trustworthiness dimensions bring to the organisation.

4.2.4. Phase 4: Vulnerability Assessment

This phase focuses on identifying potential weaknesses in the AI system, including technical vulnerabilities (such as software flaws, inadequate data governance, and network weaknesses) that could be exploited by attackers. The AI_TAF framework in this phase estimates the vulnerability level of each AI asset to each of its threats based on the missing controls; the proposed scale is based on the percentage of controls in place versus the total available controls (Table 7).

Available AI controls can be found in various knowledge databases (DB), e.g., OWASP AI [45], ENISA [42], NIST AI Risk Management Framework (AI RMF 1.0) [2], or by using AI assessment tools (see Table 1). These databases offer catalogues of AI-related threats, vulnerabilities, and controls that assessors can use as references. Technical vulnerabilities can also be identified using AI assessment tools, penetration testing, vulnerability scans, and social engineering evaluations.

The vulnerability level can be adjusted according to the criticality of the AI system and the organisation's "risk appetite." For example, for AI systems with a Very High criticality level, the vulnerability level could be Very High if a significant majority (less than 80–90%) of controls are implemented.

Table 7. Vulnerability level.

Vulnerability Level	Means
(VH) = 5	None (0%) of the available controls have been implemented to prevent the threat from being exploited.
(H) = 4	Very few (<20–40 %) of the available controls have been applied to guard against the exploitation of the threat.
(S) = 3	Few (<40–60%) of the available controls have been put in place to protect against the threat.
(M) = 2	Many (>60–80%) of the available controls have been implemented to reduce the chance of the threat being exploited.
(L) = 1	Most (>80–99%) of the available controls have been activated to prevent the threat from being exploited.

Output of Phase 4:

Vulnerability levels are recorded for each AI asset for each of its AI threats in this lifecycle under assessment.

4.2.5. Phase 5: Risk Assessment

Risk assessment in AI_TAF synthesises the findings of the previous phases. It involves assigning risk levels using:

$$\textbf{Risk (R)} = \text{Threat (T)} \times \text{Vulnerability (V)} \times \text{Impact (I)} \tag{1}$$

for each AI asset and threat under assessment using the scales in the following table (Table 8):

Table 8. Risk score calculator.

Risk Score Range	Risk Level (R)
76–125	(VH)
51–75	(H)
25–50	(M)
1–24	(L)

It is important to note that the resulting risk value is a composite ordinal index used for prioritisation and relative severity comparison. The score does not imply a probabilistic or metric interpretation (e.g., “Risk = 80” does not imply 80% likelihood). Rather, it provides a structured way to order AI assets by their estimated trustworthiness threat profile based on lifecycle phase and contextual attributes.

Although different combinations of threat, impact, and vulnerability can produce the same composite score, the AI_TAF framework retains the value of each component during reporting to preserve transparency. For example, an asset with (T = 5, V = 4, I = 1) reflects a high likelihood but minimal consequence, whereas (T = 1, V = 4, I = 5) reflects a high consequence but low likelihood. Risk owners are advised to interpret the results in conjunction with the individual component levels before mitigation.

This Risk level (R) will be refined if the estimated AI maturity level (tAIP) of the asset owner team is considered. The **refined risk level (fR)** is derived by adjusting the initial risk value (R), which is calculated for each AI asset in relation to each identified threat.

If the AI maturity level (tAIP) of the team responsible for a specific AI asset has been determined, fR can be recalculated as follows:

$$\mathbf{fR} = \begin{cases} R-1, & tAIP >= Medium\ (M) \\ R+1, & tAIP < Medium\ (M) \end{cases} \tag{2}$$

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

The use of a "−1 level" adjustment to the risk index based on the AI team's tAIP score is implemented as a heuristic correction. We acknowledge that Likert scales are ordinal and that the distances between levels (e.g., from 'high' to 'moderate') are not guaranteed to be equidistant. Therefore, subtraction should not be interpreted as a mathematical operation but as a proportional reduction in relative risk—sufficient for comparative prioritisation rather than absolute scoring.

Furthermore, if the sophistication level of the potential adversary (tA) is also determined, the proposed calculation can be further refined by incorporating this factor into the **final risk level (FR)** as follows:

$$\mathbf{FR} = \begin{cases} \mathrm{R}, & \mathrm{tA} = \mathrm{tAIP} \\ \mathrm{R}-1, & \mathrm{tAIP} > \mathrm{tA} \\ \mathrm{R}+1, & \mathrm{tAIP} < \mathrm{tA} \end{cases} \tag{3}$$

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

4.2.6. Phase 6: Risk Management

The final phase addresses how to respond to the identified risks. AI_TAF offers a layered approach:

- Proposal of Social controls to enhance the maturity level of AI teams using Table 2 according to the team's maturity level.
- Technical controls (that will complement the implemented ones) will be selected from the existing OWASP and ENISA databases or those proposed by the AI assessment tools in Table 1.

Additionally, the framework supports dynamic feedback loops, allowing teams to revisit earlier phases as systems evolve and new threats emerge. AI_TAF encourages the documentation of lessons learned and integrates them into organisational memory, enhancing maturity over time.

Phase 6 Output:

A documented overview of the proposed controls.

4.2.7. Example for Applying the AI-TAF

AI system under assessment: a simple AI-based chatbot designed to provide customer support for an e-commerce platform hosted by a small e-shop. This chatbot responds to customer queries, processes orders, and provides troubleshooting instructions.

Phase 1—Initialisation: The e-commerce platform is hosted and operated in a small shop and is not a critical AI system; according to Table 4, the criticality level is low (L). There is no need for an asset model since there is only one AI asset, i.e., the chatbot.

The trustworthiness dimensions that are relevant and interesting to the owners of this AI system are:

D1. Robustness—Ensuring that the chatbot functions correctly under different conditions.

D2. Fairness & Bias—Avoiding discrimination in responses.

The assessment will occur during the operation phase of the AI system lifecycle. The chatbot is operated only by one administrator (owner of the asset), and her trustworthiness maturity level has been reported (tAIP) = H. From previous incidents that conducted criminal investigations, the identified attackers had a sophistication level of tA = VH.

The four controls that have already been implemented are data validation procedures, data policy in place, API security, and encryption of sensitive data.

The risk appetite decided by the management of the shop is that they will treat 100% only those risks with risk levels FR = VH; they will postpone the treatment of the risks with levels FR = H until next year; they will absorb all other risks.

Phase 2—Threat Assessment: Using the OWASP AI Exchange [21], it was found that threats related to robustness, fairness, and bias and the controls appropriate to mitigate these threats are (Table 9):

Table 9. AI Security Threats and 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 sanitisation, 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, and access control for training data. **Efficient data management policy**
Model Evasion	Crafting inputs to bypass AI-based security mechanisms like spam filters.	Use of tools Model hardening, behavioural analytics, and continuous security assessments.
Model Extraction	Stealing AI models by making repeated queries to approximate their behaviour.	Rate limiting, **API security**, encrypted queries, watermarking models, and 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 inherit 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 causes slowdowns or crashes.	Rate limiting, request throttling, anomaly detection, and server-side protections.

The four controls that have already been implemented (as reported in the initialisation phase) are the ones in bold, namely data validation, data policy in place, and API security; they also apply encryption to sensitive data for the threats of data poisoning, model extraction, and model inversion respectively. Since the AI system is not critical, the risk assessor used Table 5 and provided the following threat assessment for the considered threats to the chatbot under assessment (Table 10):

Table 10. Threat levels for AI assets (Chatbot).

Threats to the AI Asset (Chatbot)	Threat Level
Manipulation	VH
Data Poisoning	VH
Model Evasion	H
Model Extraction	M
Model Inversion	L
Inherited bias from biased data	L
Overloading	M

Phase 3 Impact Assessment: The threats relative to the chatbot impact the two dimensions (D1 and D2) of trustworthiness as follows, according to Table 11:

Table 11. Impact of AI threats across dimensions D1 and D2.

Threats	Impact Dimensions (D1 and D2)
Manipulation	VH
Data Poisoning	H
Model Evasion	M
Model Extraction	H
Model Inversion	H
Inherited bias from bias data	H
Overloading	S

Phase 4: Vulnerability Assessment: The four controls that have already been implemented (as reported in the initialisation phase) partially mitigate data poisoning, model extraction, and model inversion threats.

Based on the OWASP controls identified in Phase 2, the assessor used Table 7 and reported that the vulnerability level of the chatbot to the identified threats are (Table 12):

Table 12. Vulnerability assessment of the AI asset against identified threats.

Threats	Vulnerability Level of the AI Asset to the Threats
Manipulation	VH: None of the available controls, as shown in Phase 2, have been applied to guard against the exploitation of the threat.
Data Poisoning	S: Few (<40–60%) of the available controls, as shown in Phase 2, have been applied to guard against the exploitation of the threat.
Model Evasion	VH: None of the available controls, as shown in Phase 2, have been applied to guard against the exploitation of the threat.
Model Extraction	H: Very few (<20–40 %) of the available controls, as shown in Phase 2, have been applied to guard against the exploitation of the threat.
Model Inversion	H: Very few (<20–40 %) of the available controls, as shown in Phase 2, have been applied to guard against the exploitation of the threat.
Inherited bias from bias data	VH: None of the available controls, as shown in Phase 2, have been applied to guard against the exploitation of the threat.
Overloading	VH: None of the available controls, as shown in Phase 2, have been applied to guard against the exploitation of the threat.

Phase 5: Risk Assessment: Using Table 8 and the conditional equations for the refined (fR) and final risk levels (FR), we conclude (Table 13):

Table 13. Risk assessment summary.

Threats	Threat Level	Vulnerability Level	Impact Level	R	tAIP	fR	tA	FR
Manipulation	VH (5)	VH(5)	VH (5)	VH (125)	H	H	VH	VH
Data Poisoning	VH (5)	H (4)	H (4)	VH (80)	H	H	VH	VH
Model Evasion	H(4)	S (3)	M (2)	L(24)	H	VL	VH	M
Model Extraction	M (2)	H (4)	H (4)	VH (32)	H	H	VH	VH
Model Inversion	L (1)	H (4)	H (4)	L(16)	H	VL	VH	M
Inherited bias from bias data	L (1)	VH (5)	H (4)	L(20)	H	VL	VH	M
Overloading	M (2)	VH (5)	S (3)	M (30)	H	VL	VH	S

The above table reveals how the risk calculation depends on human element consideration.

5. Further Research

Although AI_TAF provides a robust structure, several areas remain open for further exploration, and the authors continue their research in these areas.

- Quantitative trust metrics: Developing standardised, industry-wide benchmarks for trust measurements;
- Automated maturity assessment tools: AI-driven diagnostic privacy-aware tools are needed to assess team readiness without extensive manual work;
- Integration with legal frameworks: The dynamic nature of AI regulation (e.g., EU AI Act updates) requires the framework to evolve;
- Real-world piloting: Future work should focus on deploying AI_TAF in live projects across various domains to collect evidence-based feedback;
- Interdisciplinary collaboration models: Understanding how to best facilitate cooperation between developers, ethicists, legal teams, and domain experts remains an ongoing challenge.

To illustrate this, the figure below (Figure 8) highlights the key areas where these challenges can be addressed in the development of AI_TAF.

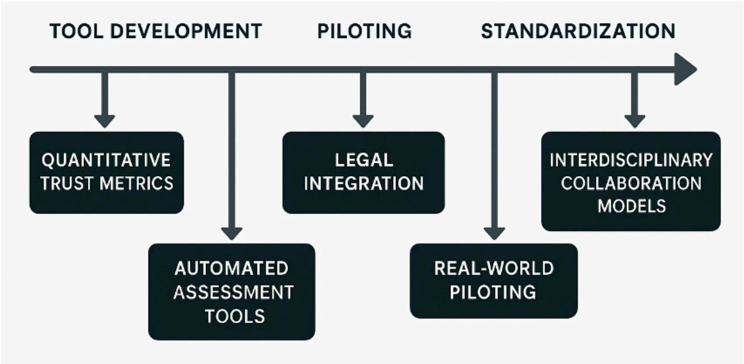

TOOL DEVELOPMENT
PILOTING
STANDARDIZATION
QUANTITATIVE TRUST METRICS
AUTOMATED ASSESSMENT TOOLS
LEGAL INTEGRATION
REAL-WORLD PILOTING
INTERDISCIPLINARY COLLABORATION MODELS

Figure 8. Research roadmap.

Extensive workshops with AI teams are needed to finalise and reach a consensus on the proposed scales in Table 3 and provide additional targeted, well-accepted social measures for enhancing the trustworthiness maturity of the AI teams.

Furthermore, the case study presented in this work—an AI-enabled chatbot in an e-commerce context—was selected to illustrate the full application of the AI_TAF methodology in a controlled and easily interpretable setting. However, we acknowledge that this example does not reflect the complexity or criticality of domains such as healthcare, finance, and legal systems. In future work, the framework will be applied to high-risk AI systems where domain-specific trust dimensions (e.g., regulatory oversight and real-time safety constraints) become prominent. Additionally, comparative evaluations will be conducted to benchmark AI_TAF against other trust frameworks in order to assess its practical effectiveness and adaptability across diverse sectors.

6. Conclusions

Ongoing advancements in Artificial Intelligence necessitate a careful, multidisciplinary approach to ensure that these technologies remain trustworthy and aligned with societal and organisational expectations. The AI_TAF addresses this need by integrating expertise from diverse domains, such as AI governance, cybersecurity, ethics, law, and risk management.

AI_TAF is a stepwise approach for evaluating and mitigating trustworthiness risks in AI systems. It can be used repeatedly in each phase (design, development, and deployment) of its lifecycle. It systematically assesses the threats and vulnerabilities of AI system assets in the specific lifecycle phase under assessment against key trustworthiness dimensions. By identifying the potential consequences of the dimensions, it estimates risks and recommends tailored mitigation strategies. AI_TAF assesses the risks of each AI system asset, considering the criticality of the AI system and the maturity of the AI teams that own (develop/design/use/interact) the asset in the lifecycle phase when the assessment takes place. By emphasising human-centric evaluation and incorporating continuous stakeholder feedback, the framework remains adaptable to real-world applications, evolves over time, and builds the necessary capabilities among individuals to oversee AI systems in alignment with the requirements of the EU AI Act. Moving forward, further validation and updates to the framework will continue to strengthen its relevance and applicability in the ever-evolving AI landscape.

Author Contributions: All authors contributed equally to this manuscript and take full responsibility of it. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Commission, Grant number 101135932.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The authors would like to acknowledge the financial support provided for the following project: The 'Fostering Artificial Intelligence Trust for Humans towards the Optimisation of Trustworthiness through Large-scale Pilots in Critical Domains' (FAITH) project, which has received funding from the European Union's Horizon Program under grant agreement No. 101135932, and all partners in this project for their contributions. The views expressed in this paper, however, represent only the views of the authors and not those of the European Commission or the partners in the abovementioned project. Finally, the authors declare that there are no conflicts of interest, including financial conflicts that could be perceived as potential conflicts.

Conflicts of Interest: Eleni Seralidou, Kitty Kioskli, Theofanis Fotis and Nineta Polemi are employed by trustilio B.V.

References

1. OECD. Governing with Artificial Intelligence: Are Governments Ready? OECD Artificial Intelligence Papers. June 2024. No 20. Available online: https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/06/governing-with-artificial-intelligence_f0e316f5/26324bc2-en.pdf (accessed on 1 May 2025).
2. National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*; (NIST AI 100-1); National Institute of Standards and Technology: Gaithersburg, MD, USA, 2023. Available online: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf (accessed on 1 May 2025).

3. European Commission. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (COM/2021/206 Final). 2021. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (accessed on 1 May 2025).
4. European Commission. Artificial Intelligence Act: Regulation of AI in the European Union. 2024. Available online: https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence (accessed on 1 May 2025).
5. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An ethical framework for a good AI society. *Minds Mach.* **2018**, *28*, 689–707. [CrossRef] [PubMed]
6. European Parliament & Council. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Official Journal of the European Union. 2024. Available online: https://eur-lex.europa.eu (accessed on 1 May 2025).
7. Stettinger, G.; Weissensteiner, P.; Khastgir, S. Trustworthiness Assurance Assessment for High-Risk AI-Based Systems. *IEEE Access* **2024**, *12*, 22718–22745. [CrossRef]
8. Jobin, A.; Ienca, M.; Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **2019**, *1*, 389–399. [CrossRef]
9. Kioskli, K.; Polemi, N. Estimating attackers' profiles results in more realistic vulnerability severity scores. In Proceedings of the 13th International Conference on Applied Human factors and Ergonomics (AHFE2022), New York, NY, USA, 24–28 July 2022; Springer: Berlin/Heidelberg, Germany; Elsevier: Amsterdam, The Netherlands; CRC: Boca Raton, FL, USA, 2022; Volume 53, pp. 138–150.
10. Kioskli, K.; Fotis, T.; Nifakos, S.; Mouratidis, H. The Importance of conceptualising the human-centric approach in maintaining and promoting cybersecurity-hygiene in healthcare 4.0. *Appl. Sci.* **2023**, *13*, 3410. [CrossRef]
11. Kioskli, K.; Seralidou, E.; Polemi, N. A Practical Human-Centric Risk Management (HRM) Methodology. *Electronics* **2025**, *14*, 486. [CrossRef]
12. Hagendorff, T. The ethics of AI ethics: An evaluation of guidelines. *Minds Mach.* **2020**, *30*, 99–120. [CrossRef]
13. Mökander, J.; Floridi, L. Ethics-based auditing of automated decision-making systems. *AI Soc.* **2021**, *36*, 511–525.
14. *ISO/IEC 23894:2023*; Information Technology—Artificial Intelligence—Guidance on Risk Management. International Organization for Standardization (ISO): Geneva, Switzerland, 2023. Available online: https://www.iso.org/standard/81228.html (accessed on 1 May 2025).
15. *ISO/IEC 27090:2018*; Information Technology—Security Techniques—Information Security Management Guidelines for Smart Cities. International Organization for Standardization (ISO): Geneva, Switzerland, 2018. Available online: https://www.iso.org/standard/73906.html (accessed on 1 May 2025).
16. *ISO/IEC 27091:2015*; Information Technology—Security Techniques—Guidance for Incident Sharing. International Organization for Standardization (ISO): Geneva, Switzerland, 2015. Available online: https://www.iso.org/standard/60803.html (accessed on 1 May 2025).
17. *ISO/IEC 5338:2023*; Information Technology—Artificial Intelligence—AI System Life Cycle Processes. International Organization for Standardization (ISO): Geneva, Switzerland, 2023. Available online: https://www.iso.org/standard/81229.html (accessed on 1 May 2025).
18. European Telecommunications Standards Institute. *TC SAI Activity Report 2023*; ETSI: Sophia Antipolis, France, 2023; Available online: https://www.etsi.org/committee-activity/activity-report-sai (accessed on 1 May 2025).
19. *ISO/IEC 42001:2023*; Information Technology—Artificial Intelligence—Management System. International Organization for Standardization (ISO): Geneva, Switzerland, 2023. Available online: https://www.iso.org/standard/81230.html (accessed on 1 May 2025).
20. European Union Agency for Cybersecurity (ENISA). Framework for AI cybersecurity practices (FAICP). 2023. Available online: https://www.enisa.europa.eu/publications/multilayer-framework-for-good-cybersecurity-practices-for-ai (accessed on 1 May 2025).
21. OWASP Foundation. OWASP AI Exchange: AI Security & Privacy Guide. 2025. Available online: https://owaspai.org/ (accessed on 1 May 2025).
22. Polemi, N.; Praca, I.; Kioskli, K.; Becue, A. Challenges and Efforts in Managing AI Trustworthiness Risks: A state of knowledge. *Front. Big Data* **2024**, *7*, 1381163. [CrossRef] [PubMed]
23. Kioskli, K.; Ramfos, A.; Taylor, S.; Maglaras, L.; Lugo, R. Optimizing AI System Security: An Ecosystem Recommendation to Socio-Technical Risk Management. In Proceedings of the 15th International Conference on Applied Human Factors and Ergonomics (AHFE), Nice, France, 22–27 July 2024; AHFE Open Access Series. Volume 159. [CrossRef]
24. *AI Fairness 360 (AIF360)*; IBM Research: Armonk, NY, USA, 2018. Available online: https://github.com/Trusted-AI/AIF360 (accessed on 10 May 2025).
25. *AI Explainability 360 (AIX360)*; IBM Research: Armonk, NY, USA, 2019. Available online: https://github.com/Trusted-AI/AIX360 (accessed on 10 May 2025).

26. *Adversarial Robustness Toolbox (ART)*; Linux Foundation AI & Data; IBM Research: Armonk, NY, USA, 2021. Available online: https://github.com/Trusted-AI/adversarial-robustness-toolbox (accessed on 10 May 2025).
27. Fairlearn Contributors. *Fairlearn*; Microsoft: Redmond, WA, USA, 2020. Available online: https://github.com/fairlearn/fairlearn (accessed on 10 May 2025).
28. Google PAIR. *What-If Tool*; Google: Mountain View, CA, USA, 2019. Available online: https://github.com/PAIR-code/what-if-tool (accessed on 10 May 2025).
29. Deepchecks, Inc. *DeepChecks*; Deepchecks: Tel Aviv, Israel, 2022. Available online: https://github.com/deepchecks/deepchecks (accessed on 10 May 2025).
30. TensorFlow. *Model Card Toolkit*; TensorFlow: Mountain View, CA, USA, 2022. Available online: https://github.com/tensorflow/model-card-toolkit (accessed on 10 May 2025).
31. Kioskli, K.; Polemi, N. A socio-technical approach to cyber risk assessment. *Int. J. Electr. Comput. Eng.* **2020**, *14*, 305–309.
32. European Commission. *Ethics Guidelines for Trustworthy AI*; High-Level Expert Group on Artificial Intelligence; European Commission: Brussels, Belgium, 2020. Available online: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed on 1 May 2025).
33. Yu, T.; Alì, T. Governance of artificial intelligence: A risk and guideline-based approach. *AI Soc.* **2019**, *34*, 134–143.
34. OECD. OECD Framework for the Classification of AI Systems. In *OECD Digital Economy Papers*; No. 331; OECD Publishing: Paris, France, 2020. [CrossRef]
35. Florea, N.V.; Croitoru, G. The Impact of Artificial Intelligence on Communication Dynamics and Performance in Organizational Leadership. *Adm. Sci.* **2025**, *15*, 33. [CrossRef]
36. Bashkirova, A.; Krpan, D. Confirmation bias in AI-assisted decision-making: AI triage recommendations congruent with expert judgments increase psychologist trust and recommendation acceptance. *Comput. Hum. Behav. Artif. Hum.* **2024**, *2*, 100066. [CrossRef]
37. Joksimovic, S.; Ifenthaler, D.; Marrone, R.; De Laat, M.; Siemens, G. Opportunities of artificial intelligence for supporting complex problem-solving: Findings from a scoping review. *Comput. Educ. Artif. Intell.* **2023**, *4*, 100138. [CrossRef]
38. Tomazevic, N.; Murko, E.; Aristovnik, A. Organizational Enablers of Artificial Intelligence Adoption in Public Institutions: A Systematic Literature Review. *Cent. Eur. Public Adm. Rev.* **2024**, *22*, 109. [CrossRef]
39. Reuel, A.; Connolly, P.; Jafari Meimandi, K.; Tewari, S.; Wiatrak, J.; Venkatesh, D.; Kochenderfer, M. Responsible AI in the global context: Maturity model and survey. *arXiv* **2024**. [CrossRef]
40. MITRE Corporation. MITRE ATT&CK®: Groups. Available online: https://attack.mitre.org/groups/ (accessed on 10 May 2025).
41. Kioskli, K.; Polemi, D. Measuring Psychosocial and Behavioural Factors Improves Attack Potential Estimates. In Proceedings of the 2020 15th International Conference for Internet Technology and Secured Transactions (ICITST), London, UK, 8–10 December 2020; pp. 1–4. [CrossRef]
42. European Union Agency for Cybersecurity. *Artificial Intelligence Threat Landscape*; European Union Agency for Cybersecurity: Athens, Greece, 2020. Available online: https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges (accessed on 1 May 2025).
43. *ISO/IEC 27005:2022*; Information Security, Cybersecurity and Privacy Protection–Guidance on Managing Information Security risks. International Organization for Standardization (ISO): Geneva, Switzerland; International Electrotechnical Commission: Geneva, Switzerland, 2022. Available online: https://www.iso.org/standard/80585.html (accessed on 1 May 2025).
44. *ISO/IEC 27001:2022*; Information Security, Cybersecurity and Privacy Protection–Information Security Management Systems–Requirements. International Organization for Standardization (ISO): Geneva, Switzerland; International Electrotechnical Commission: Geneva, Switzerland, 2022. Available online: https://www.iso.org/standard/82875.html (accessed on 1 May 2025).
45. OWASP Foundation. Threats Through Use. OWASP AI Exchange. Available online: https://owaspai.org/docs/2_threats_through_use/ (accessed on 10 May 2025).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.