Ref. Ares(2026)3410542 - 31/03/2026

D2.2

FAITH Methodological Framework and Requirements Analysis v2

Related Work Package	**WP2 – FAITH AI Framework for trustworthiness & coordination among Large Scale Pilots (LSPs)**
Related Task	Task 2.2. - FAITH Trustworthiness Assessment Framework Version 2 (FAITH_TAF-v2)
Lead Beneficiary	UOS
Contributing Beneficiaries	FORTH, trustilio, CNR, ATC, ICCS, KU LEUVEN
Document version	**v.6 (FINAL)**
Deliverable Type	R-Document, Report
Distribution level	PU-Public
Contractual Date of Delivery	31/03/2026
Amendment Proposal	
Actual Date of Delivery	31/03/2026

Contributors	**N. Fair (UOS), Samuel Senior (UOS), Betul Gokkaya (UOS)** **N. Tachos (FORTH), B. Pezoulas (FORTH), G. Kalliatakis (FORTH), K. Vardakas (FORTH)** **E. Tsalapati (ATC)** **G. Mentzas (ICCS), D. Apostolou (ICCS), E. Anagnostopoulou (ICCS), N. Grammatikos (ICCS)** **Nineta Polemi (trustilio), Eleni Seralidou (trustilio), T. Fotis (trustilio), P. Kontari (trustilio), K. Kioskli (trustilio)** **Sara Colantonio (CNR) Claudia Caudai (CNR) Giulio del Corso (CNR) Davide Moroni (CNR) Maria Antonietta Pascali (CNR)** **Panagiota Papageorgiou (KUL), A. M. Corrêa (KUL)**
Reviewers	Asbjørn Følstad (SINTEF), 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
	Proposed TOC for D2.2.	28/11/2025	Trustilio & UoS
	Finalised ToC	3/12/2025	All
V1	Contributions	31/12/2025	T2.1 partners
V2	Consolidation to Draft-v1	26/2/2026	D2.2 leader (UOS)
V3	Internal WP Review	3/03/2026	Core WP2 Team
V4	External FAITH Review	10/03/2026	SINTEF and AOA
V5	Submission to Coordinator	27/03/2026	WP2 leader (UOS)
V6 (Final)	D2.2 Submission	31/03/2026	Coordinator (FORTH)

Statement of Originality

This deliverable contains original unpublished work except where clearly indicated otherwise.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

IP Management Statement

The FAITH TAF-v2, as a methodology, will be released openly under a Creative Commons licence, while all FAITH technical implementations of the FAITH TAF-v2 as software tools and services will remain the IP of one or more of the developers of that tool or service.

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 presents the FAITH AI Trustworthiness Assessment Framework version 2 (TAF-v2), which constitutes a significant advance from version 1. While maintaining a fundamental consistency with key aspects of version 1, through a process of collaboration, TAF-v2 has been enhanced to improve its applicability to real-world AI systems through the introduction of targeted extensions.

The most important extension has been the move from assessing the trustworthiness of isolated assets (TAF-v1) to viewing trustworthiness assessment as a process of understanding the whole ecosystem of linked assets that includes other linked software and processes, hardware and network infrastructure, physical locations and jurisdictions, as well as the core data, model and human assets. This is consistent with the TAF-v2 view that the trustworthiness of an AI system can only be effectively established through a detailed process of in-context system-level risk assessment.

To enable such an approach, and in an enhancement from TAF-v1, TAF-v2 explicitly takes into account cascading and propagating risks across assets, systems and wider social contexts. It suggests that the Impact of these cascading risks and the presence (or not) of existing Controls needs to be identified and modelled in order to allow deep context customisation, thereby making the risk assessment reflect the specific context in which the AI system is situated. It is possible to support this through the integration of the outputs of the TAF-v1 in-isolation assessments as inputs to the whole-system modelling espoused in TAF-v2, thus integrating both the TAF versions and the different layers of risk assessment (asset-level and system-level).

TAF-v2 also presents a more detailed understanding and classification of AI Model types and the specific Threats and Controls that are associated with them, as these reflect the AI systems deployed in the FAITH Large Scale Pilots as well as representing an emerging domain. It also introduces a move from single-dimension trustworthiness optimisation to balanced, evidence-based trade-off management between and across trustworthiness attributes and operational constraints. It also updates the policy, legislation, standards and metrics landscape with the most recent AI-specific developments.

Finally, the individual steps associated with each of the 6 TAF-v1 Phases have been refined and redefined and the need to document the actions taken at each step added. A formal validation of TAF-v2 with different groups of stakeholders (developers, risk assessors and non-technical

users) has been planned and will be implemented later in the project, and TAF-v2 will be formally validated with three groups of stakeholders (developers, risk assessors and non-technical users) operationalized in the technical iterative development process for the second versions of the FAITH Tools that will be used in the second pilot phase.

1. Table of Contents

Contents

1 Introduction 9
1.1 Overview of Key Updates from the FAITH AI Trustworthiness Assessment Framework (TAF) Version 1 10
1.2 Scope and Methodology 11
2 Updated Legislation, Policies, Standards & Metrics 12
2.1 Legislation & Policies 12
2.1.1 NIS II Directive 12
2.1.2 Cyber Resilience Act (CRA) 12
2.1.3 Digital Services Act (DSA) 13
2.1.4 Digital Markets Act (DMA) 14
2.1.5 AI HLEG Principles 14
2.2 Evolving Regulatory Landscape 15
2.2.1 The Digital Omnibus Regulation Proposal 16
2.3 AI Risk Assessment Standards 17
2.4 Standards for Agentic AI systems 20
2.5 AI Risk Frameworks 22
2.6 Updates to AI Metrics 23
2.6.1 Metrics for the TWA: Fair with Harmful Bias Managed - Data Bias Detection 23
2.6.2 Metrics for the TWA: Fair with Harmful Bias Managed - AI Model Bias Detection 24
2.6.3 Metrics for the TWA: Privacy-Enhanced - Data Privacy Assessment 25
2.6.4 Metrics for the TWA: Explainable and Interpretable - Explainability Evaluation 26
2.6.5 Metrics for the TWA: Secure and Resilient - Adversarial Robustness and Security 27
2.6.6 Metrics for Efficiency and Sustainability (Green AI) 28
2.6.7 Metrics for Uncertainty and Calibration 29
3 An In-context Risk Assessment-Based Approach to AI Trustworthiness 31
3.1 Interconnected and Interdependent Systems 31
3.2 Importance of Impact in In-context Risk Assessment 33

3.3 Humans in the System 34
3.4 Cascading Threats 35
3.5 Classifying AI Models and Systems 39
3.5.1 Classification by Model or System Type 40
3.5.2 Classification by Learning Paradigm 43
3.5.3 Classification by Function 45
3.6 AI-specific Risks 47
3.6.1 LLM-specific Attacks, Threats, Risks & Vulnerabilities 48
3.6.2 Threats Arising from Linked LLM Assets 50
3.6.3 Cascading Threats and Systemic Risk 51
3.6.4 Agentic AI-specific Attacks, Threats, Risks and Vulnerabilities 52
3.7 Trade-offs and Holistic Management 56
3.7.1 Interdependencies Among Dimensions of Trustworthiness in AI Systems 56
4 The FAITH AI-TAF - v2 59
4.1 FAITH AI Trustworthiness Assessment Framework-v2 Scope, Principles, and Uplifted Assumptions 59
4.1.1 Phase 1: Cartography 61
4.1.2 Phase 2: Threat Analysis 71
4.1.3 Phase 3: Impact Assessment 74
4.1.4 Phase 4: Vulnerability Analysis 75
4.1.5 Phase 5: Risk Analysis 85
4.1.6 Phase 6: Countermeasures 88
5 Implementation and Validation of the FAITH AI TAF-v2 91
5.1 TAF-v2 Validation Plan 91
5.2 TAF-v2 Implementation in FAITH Tools 91
6 Conclusions 93

List of Abbreviations

Abbreviation	Explanation
AI	Artificial Intelligence
AI HLEG	EU High-Level Expert Group on Artificial Intelligence
CEN	European Committee for Standardisation
CENELEC	European Committee for Electrotechnical Standardisation
CRA	Cyber Resilience Act
DMA	Digital Markets Act
DSA	Digital Services Act
ENISA	European Union Agency for Cybersecurity (originally European Network and Information Security Agency)
ETSI	European Telecommunications Standards Institute
EU	European Union
GDPR	General Data Protection Regulation
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
ISO	International Organisation for Standardisation
LLM	Large Language Model
NIST AI RMF	National Institute of Standards and Technology Artificial Intelligence Risk Management Framework
OWASP	Open Worldwide Application Security Project
TAF-v1	The first version of the FAITH AI Trustworthiness Assessment Framework
TAF-v2	The second version of the FAITH AI Trustworthiness Assessment Framework
TWA	Trustworthiness Attribute
SSM	The System Security Modeller tool Artificial Intelligence Risk Management Framework

List of Tables

Table 1: AI Model Bias Metrics 24
Table 2: Data Privacy Metrics 25
Table 3: Explainability Metrics 26
Table 4: Adversarial Robustness and Security Metrics 28
Table 5: Efficiency and Sustainability Metrics 29
Table 6: Calibration Metrics 30
Table 7: AI Model Classification by Interconnectedness 40
Table 8: ENISA & NIST Asset-to-Lifecycle Mapping 67
Table 9: Example AI-specific Technical Vulnerabilities, Controls and Limitations across AI Lifecycle Stages 77
Table 10: AI-specific Human/Social Vulnerabilities, Controls and Limitations 81

List of Figures

Figure 1: The FAITH AI Trustworthiness Assessment Framework version 2 11
Figure 2: The FAITH AI Trustworthiness Assessment Framework version 2 (repeated) 61

1 Introduction

This deliverable presents the second version of the FAITH Trustworthiness Assessment Framework (FAITH AI TAF-v2), building upon the methodological foundations established in the first version of the TAF reported on in Deliverable D2.1. The objective of this updated version is not to completely redefine the framework, but to enhance its applicability to real-world AI systems through targeted extensions. There are a range of challenges facing risk assessment used for trustworthiness that were not fully met by TAF-v1 and that, as a result, TAF-v2 addresses.

As with many traditional AI assessment frameworks, TAF-v1 takes an in-isolation approach, which does not consider the AI asset as situated within a socio-technical ecosystem that includes its complex network of relationships with other software and processes, hardware and network infrastructure, people, places and jurisdictions. Nor does an in-isolation assessment account for the specifics of the real-world context in which that AI system is located. Nor does it account for cascading effects as Threats, Consequences and Impacts propagate and multiply across assets, systems and wider ecosystems. Additionally, in-isolation assessment is also only capable of providing recommendations for asset-level Controls that serve more as a patch fix for an asset than as a system-wide mitigation strategy. Together, this means in-isolation assessment can result in incomplete risk identification, unrealistic trustworthiness assessment and sub-optimal mitigation implementation.

To address this, TAF-v2 has been extended throughout to integrate in-isolation asset-level assessment with in-context system-level assessment, where asset level outcomes can help fine-tune system level modelling. It introduces the notion of in-isolation assessment outputs as a fine-tuning in-context inputs, context-specific system modelling, cascading effects, and Control Strategies (as opposed to individual Controls).

In a further extension of the understanding of the interconnectedness of AI systems in the real, TAF-v2 introduces the concepts of trade-offs, holistic risk management, and risk tolerance assessment. Not only are assets in an AI ecosystem linked and co-dependent, but so too are the trustworthiness dimensions associated with them. This leads to the need to undertake evidence-based trustworthiness trade-offs alongside trade-offs based on other factors (e.g. effort, financial and reputational costs, system performance, organisational capability…etc), where enhancing trustworthiness in one dimension must be balanced against the impact that will have on other dimensions in a process of holistic risk management. These considerations then play out in the establishment of a risk tolerance (acceptable residual risk) that aligns with the specific context in which the AI system is situated.

In addition, TAF-v2 updates the wider policy, legislation, standards and metrics landscape to incorporate changes in this domain since TAF-v1 and to explore potential future changes. It has added nuance to the classification of AI models and the specific Threats and Controls associated with them (in particular those associated with LLM and Agentic AI systems, as these are the types found in the FAITH Large Scale Pilots). It has also revisited and redefined the steps associated

with each of the 6 Phases presented in TAF-v1 and incorporated the need to document actions taken in each those steps, as summarised in Figure 1 below.

Despite this expanded scope, the foundational principles of the framework remain consistent with those introduced in TAF-v1. FAITH AI TAF-v2 retains a human-centric approach, recognising that trustworthiness is both shaped by human decision-making, oversight, and governance throughout the AI lifecycle, and dependent on human understanding. Human actors are explicitly considered as potential sources of risk, as essential contributors to effective risk mitigation (through trade-off decision-making), and as the ultimate arbiter of whether an AI system is deemed trustworthy.

In addition, TAF-v2 continues to align with established risk management and AI governance standards, in particular the principles articulated by ENISA, NIST and ISO, noting any updates and evolutions of them. The updated TAF-v2 strengthens this alignment by operationalising these principles within a socio-technical and in-context assessment framework. The foundational Layer 1 assumptions detailed in TAF-v1 (see D2.1 section 4.1) remain unchanged, except for the final assumption that excludes cascading threats – which are now included. The understanding of the three lifecycle stages (design / testing / production) also remain unchanged, although the various AI assets have now been mapped to the most relevant lifecycle stage.

As a result, the extensions and developments presented in TAF v2 offer a more holistic and operationally applicable framework for assessing trustworthiness in complex AI systems, while preserving methodological continuity with the first version.

1.1 Overview of Key Updates from the FAITH AI Trustworthiness Assessment Framework (TAF) Version 1

The key differences between FAITH AI TAF-v1 and v2 can be summarised as follows:

- A move from the assessment of isolated assets (data, model, human) to that of ecosystems of connected assets (other software and processes, hardware and network infrastructure, people, places and jurisdictions) – see Chapter 3
- An explicit accounting of cascading and propagating risks across systems – see Section 3.4
- The inclusion of context-specific fine-tuning of Impact Levels and existing Controls for all assets in a system model that allows for deep context-customisation - this is particularly enabled for key assets that have been the subject of in-isolation risk assessment (data, models, humans) by the integration of information flows from in-isolation to in-context risk assessments – see Section 4.1.3 and Section 4.1.4
- A more detailed understanding and classification of AI Model types and their still-emerging associated Threats and Controls – see Section 3.5
- The inclusion of system-level Control Strategies – see Section 4.1.4.3
- A move from single-dimension trustworthiness optimisation to balanced, evidence-based trade-off management across trustworthiness attributes and operational constraints – see Section 3.7

- Refined and redefined steps for each of the 6 TAF Phases - along with the need to document actions taken at each step (see Figure 1) – see Chapter 4

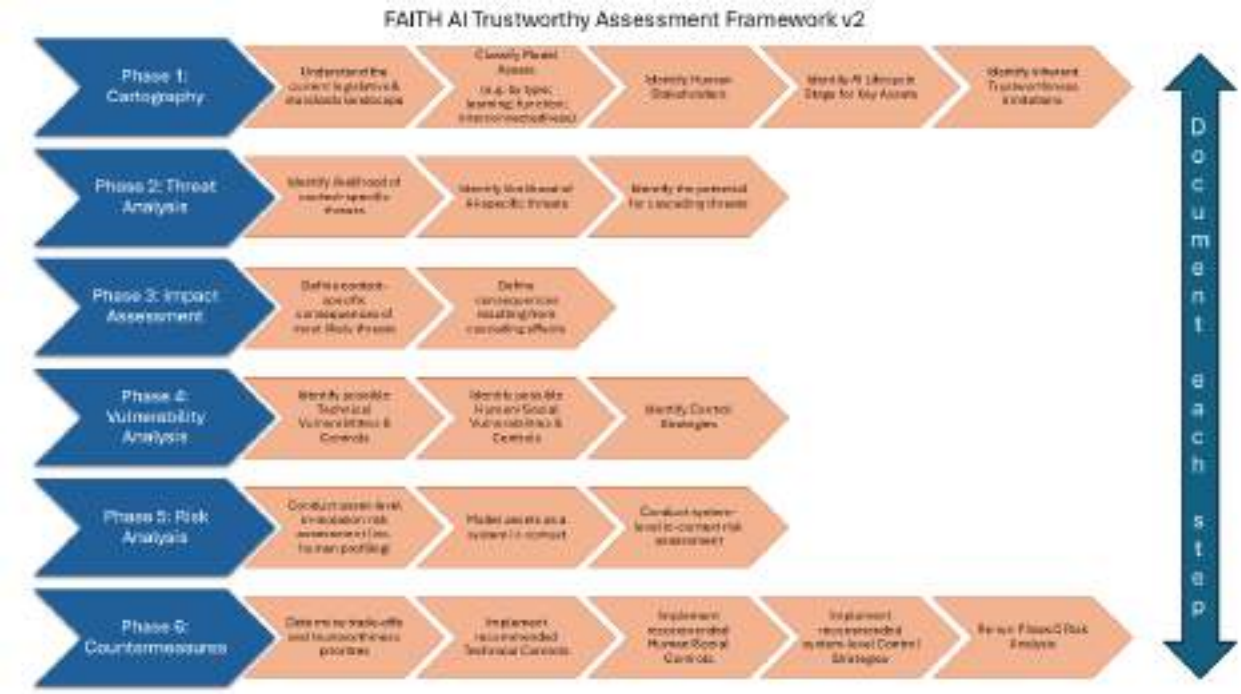


Figure 1: The FAITH AI Trustworthiness Assessment Framework version 2

1.2 Scope and Methodology

TAF-v2 is a theoretical framework designed to support risk assessors seeking to undertake a socio-technical in-context system-level approach to determining the trustworthiness of an AI system within its specific context through accurate risk assessment and the identification of appropriate mitigation actions and residual risk levels through evidence-based trade-off decision-making. Although the FAITH project has developed a series of tools that align with the TAF to support such an approach, this deliverable is not intended to present any of the technical developments of those tools, which will be detailed in D2.4 later in the project.

TAF-v2 was developed through a process of multi-disciplinary collaboration between WP partners, where updates were agreed by all contributors through a series of meetings and discussions. This was then followed by a period of joint content writing for D2.2. The deliverable document was then compiled and cleaned and the narrative structure finally established, before a detailed WP-internal review of the TAF (and D2.2) was undertaken. Thereafter, the deliverable was subject to wider FAITH review. Suggested improvement to both the TAF and the document were then implemented before final submission. A TAF-v2 validation plan was developed, which will be implemented from M30 to M36.

2 Updated Legislation, Policies, Standards & Metrics

2.1 Legislation & Policies

This section sets out the latest developments in the EU legal and policy landscape that are relevant to AI systems. Building on Deliverable D2.1 (TAF-v1), it updates with recently introduced legal and policy instruments and indicates where these may have implications for AI system risk assessment from a compliance perspective. It should be noted that compliance assessments are normally undertaken through formal Conformity Assessments, internal documentation and management processes, and external audit, rather than being part of a risk assessment process. They are included here because they do form part of the socio-technical whole-system in-context risk assessment landscape, however, the ability for FAITH tools and services to account for this is limited to an assessment of General Data Protection Regulation (GDPR) compliance only.

2.1.1 NIS II Directive

As the successor to the original NIS Directive, Directive (EU) 2022/2555[1], the NIS II Directive has expanded the scope, strengthened measures, and introduced new requirements to help protect critical infrastructures, essential services, and key sectors from cyber threats.[2] Since first reported in D2.1, NIS II entered into force in 2023 with implementation during 2024-2025.

Under NIS II, essential and important entities are required to adopt appropriate and proportionate technical, operational, and organisational measures to manage risks to the security of network and information systems (NIS II, art. 21). These measures must take into account the state of the art, relevant standards, implementation costs, and the entity's exposure to risk, and must follow an all-hazards approach, including supply chain security. By upgrading EU-wide cybersecurity risk-management and incident-reporting obligations (NIS II, art. 23) for "essential" and "important" entities, including many operators that develop or deploy AI systems, NIS II further evidences the growing legislative emphasis on in-context whole-system risk assessment. In particular, NIS II interacts with AI ACT obligations on Robustness and cybersecurity for high-risk AI, and as such raises a further potential compliance risk.

2.1.2 Cyber Resilience Act (CRA)

The EU Cyber Resilience Act is a regulation that sets mandatory, harmonised cybersecurity requirements for hardware and software products with digital elements placed on the EU market, to ensure they are securely designed, developed, updated and supported throughout

[1] Directive (EU) 2022/2555 of the European Parliament and of the Council on measures for a high common level of cybersecurity across the Union (NIS 2 Directive)

[2] ENISA, "Network and Information Systems Directive 2 (NIS2) – Awareness and Cyber Hygiene", European Union Agency for Cybersecurity, https://www.enisa.europa.eu/topics/awareness-and-cyber-hygiene/raising-awareness-campaigns/network-and-information-systems-directive-2-nis2.

their lifecycle. Since D2.1, the Cyber Resilience Act[3] was adopted in 2024 and will be applied in phases from 2026 with full application in December 2027. It mandates that products with digital elements which are made available on the market, and which are intended to be used or reasonably used in connection with a device or network, including interconnectedness with AI models, must undergo a cybersecurity risk assessment (CRA, art. 13). The cybersecurity risk assessment must consist of an analysis of cybersecurity risks considering the intended purpose and reasonably foreseeable use, the conditions of use, the length of time the product is expected to be in use, and an indication of how the cybersecurity requirements are implemented in practice (CRA, art 13(3)). The CRA imposes cybersecurity and vulnerability-management requirements throughout the life cycle of digital products, including software and potentially embedded AI components, and again evidences the legislative focus on in-context whole-system risk assessment. It compliments the AI ACT in relation to Safety and Robustness duties, as it provisions that a high-risk AI system that meets the requirements of the CRA is automatically compliant with the requirements posed by article 15 of the AI ACT (Article 12 CRA). This introduces the interesting notion of cascading compliance risks and mitigations across separate legislative instruments, which is a promising area of future research.

2.1.3 Digital Services Act (DSA)

Although not directly targeting the AI domain, since late 2022, the Digital Services Act (Regulation (EU) 2022/2065)[4] has overhauled the rules for online intermediary services such as hosting providers, online platforms, marketplaces, social networks and search engines. Its aim is to create a safe and trustworthy digital environment while preserving fundamental rights and the free movement of digital services in the single market.

It requires due-diligence duties: notice-and-action mechanisms for illegal content; “statements of reasons” for removals; complaint and redress systems; and transparency reporting about content moderation and automated tools. It bans certain targeted advertising practices (e.g. targeting children and using sensitive data), outlaws dark patterns, and requires marketplaces to verify trader identities. Very Large Online Platforms and Search Engines, many of which are now AI-driven and/or provide AI services, face the strictest obligations. They must conduct recurring systemic-risk self-assessments that explicitly cover the design and functioning of their algorithmic systems, implement mitigation measures, undergo independent audits, and grant data access to regulators and vetted researchers. The DSA requires recommender systems to be explainable in terms and conditions and to offer meaningful user controls, including non-profiling options. This emphasises the importance of on-going monitoring and repeated risk assessment. For certain specific AI systems, the DSA has the potential to introduce new compliance risks related to

[3] Regulation (EU) 2024/2847 of the European Parliament and of the Council on horizontal cybersecurity requirements for products with digital elements (Cyber Resilience Act), 2024.

[4] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)

amplification, disinformation, discrimination and other societal harms caused by AI-mediated content curation, which links with AI ACT Safety and Transparency duties.

The DSA does not directly impact the FAITH project but is included here for contextual completeness.

2.1.4 Digital Markets Act (DMA)

Th Digital Markets Act (Regulation (EU) 2022/1925)[5], whose obligations have been phasing in since 2023, targets large "gatekeeper" platforms that control core services such as app stores, search engines, social networks, operating systems and online intermediation services. It supplements competition law with an ex-ante rulebook: designated gatekeepers must not self-preference their own services; combine personal data across services without consent; lock in business users; or restrict off-platform offers, and they must enable interoperability, data portability and access to data generated via their platforms.

Again, although not directly related to AI systems, the DMA obligations do have an indirect shaping effect on AI ecosystems. It reshapes data and distribution power: gatekeepers must offer fair access to data and the interfaces needed to build and deploy competing AI models (e.g. recommendation engines, assistants, ad-tech); and they must reduce "data moats" and forced bundling. This introduces new DMA-driven risks, such as self-preferencing via AI ranking, discriminatory access to training/usage data, and unfair conditions for integrating third-party AI services.

The DMA does not directly impact the FAITH project but is included here for contextual completeness.

2.1.5 AI HLEG Principles

The Ethics Guidelines for Trustworthy AI[6], issued in 2019 by the EU High-Level Expert Group on Artificial Intelligence (AI HLEG), define "trustworthy AI" as AI that is lawful, ethical and robust throughout its lifecycle. They articulate four underlying ethical principles (respect for human autonomy, prevention of harm, fairness, explicability) and translate these into seven concrete requirements for AI systems:

- human agency and oversight
- technical robustness and safety
- privacy and data governance
- transparency
- diversity, non-discrimination and fairness
- societal and environmental well-being
- accountability

[5] https://www.eu-digital-markets-act.com/Digital_Markets_Act_Articles.html
[6] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Though non-binding, the Guidelines have strongly influenced EU policy, standardisation and corporate practice, becoming a de facto reference for governance frameworks and internal policies. In comparison with other requirements sets, those provided by the AI HLEG Principles introduce wider considerations, in particular social and environmental well-being, which focuses on the idea that AI should benefit all humans, support sustainability, avoid harming democracy and social cohesion, and consider impacts on future generations and other living beings. However, this is complex to measure and assess in an empirical fashion within a risk assessment context. Rather the companion Assessment List for Trustworthy AI (ALTAI)[7], which operationalises the seven requirements into a structured self-assessment questionnaire, provides a series of questions designed to guide respondents to identify, assess and mitigate environmental and social impact and harm. The ALTAI questionnaire has fed directly into compliance and risk-assessment practice, with some organisations choosing to use HLEG principles and ALTAI as overlays on top of legal compliance (e.g. GDPR, AI Act), to scope AI risks, define controls (human-in-the-loop, documentation, fairness checks, robustness testing) and evidence "ethical" and organisational accountability beyond minimum law.

2.2 Evolving Regulatory Landscape

It should also be noted that there has been a degree of recent criticism leveled at the EU concerning the extent of its legislation of the AI domain, which tech firms, start-ups and businesses consider as over-regulated, complex and negatively impacting competitiveness[8]. There is concern that the cumulative obligations of the AI Act, GDPR, copyright, product liability, the Data Act, the Data Governance Act, the CRA, NIS II, and DMA/DSA create an unclear, overlapping and increasingly complex framework that may raise significant barriers to innovation. The EU has responded to this, to its own Draghi Report[9] into EU competitiveness, and to divergent global legislative trends, by presenting the Digital Omnibus Regulation Proposal[10] [11]. The proposal aims to simplify and harmonise the EU legislative landscape includes a set of technical amendments to a large corpus of EU digital legislation, selected to bring immediate relief to businesses, public administrations, and citizens alike, and to stimulate competitiveness.[12]

[7] https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

[8] https://www.politico.eu/article/world-weary-europe-eu-approach-ai-new-delhi-india/

[9] https://commission.europa.eu/topics/competitiveness/draghi-report_en

[10] Proposal for a Regulation of the European Parliament and of the Council amending Regulations (EU) 2016/679, (EU) 2018/1724, (EU) 2018/1725, (EU) 2023/2854 and Directives 2002/58/EC, (EU) 2022/2555 and (EU) 2022/2557 as regards the simplification of the digital legislative framework, and repealing Regulations (EU) 2018/1807, (EU) 2019/1150, (EU) 2022/868, and Directive (EU) 2019/1024 (Digital Omnibus), 2025.

[11] Proposal for a Regulation of the European Parliament and of the Council amending Regulations (EU) 2024/1689 and (EU) 2018/1139 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI), 2025.

[12] European Commission, Digital Omnibus Regulation Proposal, Shaping Europe's digital future, 19 November 2025, available at: https://digital-strategy.ec.europa.eu/en/library/digital-omnibus-regulation-proposal.

2.2.1 The Digital Omnibus Regulation Proposal

This subsection will explore the potential impact of the Digital Omnibus Regulation Proposal on the legislative landscape surrounding the design, development and deployment of AI systems. On 19 November 2025, the European Commission presented the Digital Omnibus Regulation Proposal, which is currently subject to the ordinary legislative procedure. It focuses on simplifying data legislation, cybersecurity incident reporting obligations, and the smooth application of Artificial Intelligence Act (AI ACT)[13] [14] rules. Among many others, it proposes targeted amendments to GDPR and the AI ACT.

Among the proposed changes, a new legal basis is introduced under Article 9 GDPR for the processing of special categories of personal data in the context of the development and operation of AI systems, as defined in Article 3(1) of the AI ACT. In this context, the Commission proposes to add Article 9(5) GDPR, which sets out additional conditions for the processing of sensitive personal data. Additionally, the proposal introduces a new Article 88c GDPR, which would allow reliance on legitimate interests under Article 6(1)(f) GDPR as a legal basis for the development and operation of AI systems. Both additions provide an allowance to process personal data for the development and operation of AI systems.

In addition, in order to design and develop a future-proof, human-centric and marketable AI technology, the legal requirements of the AI ACT must be taken into consideration. The Digital Omnibus Regulation Proposal, introduces targeted amendments to the AI ACT. One of the most significant reforms concerns the timing of obligations for high-risk AI systems. As currently adopted, the AI ACT will become applicable from 2 August 2026, although some provisions are subject to earlier or later applicability. For example, Chapter II on Prohibited AI Practices and provisions on AI literacy have applied 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 will apply from 2 August 2027 (AI ACT, art. 113).

The updated Article 113, proposes a shift in the timing of the application of core requirements for high-risk AI systems. Rather than relying on a fixed implementation date, the application of obligations would be linked to the availability of harmonized standards, common specifications, or relevant Commission guidelines. Under this approach, requirements would become applicable six or twelve months after a Commission decision confirming regulatory readiness (depending on the category of the AI system) or, in the absence of such a decision, no later than December 2027 or August 2028, depending on the classification of the high-risk AI system.

Beyond the application timeline, the Digital Omnibus on AI introduces in the proposed Article 4a (which replaces current Article 10(5) AI ACT) a new explicit legal basis, allowing providers and deployers, under strict safeguards, to process special categories of personal data for bias

[13] 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), 2024.

[14] Ebers, M., "Truly Risk-Based Regulation of Artificial Intelligence: How to Implement the EU's AI Act," European Journal of Risk Regulation, 2024.

detection and correction where no alternative exists. This provision introduces a clearer legal basis for bias testing and strengthens responsible AI development while ensuring strict privacy protections.

While the Digital Omnibus Regulation Proposal has not yet been adopted, it is likely to have implications for how certain data processing activities within the full AI system lifecycle are assessed from a GDPR perspective, and introduce new compliance risks associated with certain data processing activities within the full AI system lifecycle from a GDPR perspective and with certain bias testing requirements within the amended AI ACT.

It will remain important to continue to monitor these changes to the EU regulatory landscape as it is a rapidly evolving space.

2.3 AI Risk Assessment Standards

Standards play a crucial role in the AI governance ecosystem, complementing and enabling implementation of international agreements, regulations, and industry self-governance[15]. Standards are central to ensuring the trustworthiness of AI systems across all stages of AI development and across a wide variety of AI applications[16]. They help to establish data governance frameworks, enable effective risk management, and contribute to sustainable development.

With multiple regulatory requirements to conduct risk assessments at all stages of the AI Lifecycle across a range of aspects, standards are relevant not as mandatory instruments, but as reference frameworks that inform how AI risks and risk assessment may be structured, interpreted and documented. Compliance with standards may also provide a presumption of conformity with the relevant legal requirements[17]. This section therefore identifies the standards that are most relevant for AI risk assessment and governance.

At the international level, standardisation activities relevant to AI risk assessment are primarily developed by ISO, IEC, and IEEE, while at the European level CEN, CENELEC, and ETSI contribute standards aligned with EU legislation and policy objectives (see List of Abbreviations for details). In line with EU standardisation practice, international standards are adopted where they are consistent with EU values and fundamental rights, while remaining gaps are addressed through European harmonised standards developed to support Union law[18].

Among the most significant contributions of ISO/IEC standards are ISO/IEC 24028, ISO/IEC 42001, and ISO/IEC 23894:2023 which provide important guidelines on trustworthiness and

[15] International Organization for Standardization, ISO policy brief: Harnessing international standards for responsible AI development and governance, 2023.
[16] European Commission, Joint Research Centre, AI Watch: AI Standardisation Landscape – State of Play and Link to the EC Proposal for an AI Regulatory Framework, 2021.
[17] European Commission, Standardisation of the AI Act, Shaping Europe's digital future, 2025
[18] Kilian, R., Jäck, L. and Ebel, D., "European AI Standards – Technical Standardisation and Implementation Challenges under the EU AI Act," 2024.

management of AI systems. ISO/IEC 24028 is a technical report that aims to define what "trustworthy" AI means by providing a shared vocabulary and conceptual framework. It identifies key trust dimensions—reliability, robustness, safety, security, privacy, transparency, fairness and accountability—and relates them to the AI lifecycle from design to decommissioning. It underpins other standards such as ISO/IEC 42001 (AI management systems) and ISO/IEC 23894 (AI risk management) and aims to support regulatory alignment and demonstrable trustworthy AI.

The ISO/IEC 42001 standard establishes a framework for organizations to systematise how organisations govern AI across its lifecycle. It defines policies, roles, processes and controls to ensure AI is developed and used responsibly, aligning with ethical, legal and business objectives by implementing, maintaining, and continually improving an AI management system[19].

ISO/IEC 23894:2023 provides guidance for managing AI-specific risks by extending the generic ISO 31000 risk-management framework to AI. It helps organisations systematically identify, analyse, evaluate and treat risks arising across the AI lifecycle, including data, model, operational and ethical risks. Its importance lies in translating abstract "trustworthy AI" goals into a structured, repeatable risk-management process that can integrate with governance frameworks (e.g. ISO/IEC 42001) and support compliance with emerging AI regulations, while improving safety, reliability and stakeholder trust in AI systems. Together, this stack of ISO standards help organisations manage the risks and governance challenges associated with AI systems[20].

In addition to AI-specific standards, general risk management standards provide an important foundation for structuring AI risk governance. ISO 31000:2018 establishes internationally recognised principles and processes for risk management, including risk identification, analysis, evaluation, and treatment across organisational activities. Recent guidance from the European Data Protection Supervisor (EDPS) has explicitly relied on ISO 31000 as a reference framework for assessing and managing risks arising from AI systems[21].

Beyond these core standards, a number of ISO/IEC standards focus on issues such as bias, data quality, robustness, safety, and cybersecurity in AI systems. These include standards addressing bias identification and mitigation (such as ISO/IEC TR 24027 and ISO/IEC TS 12791), robustness and resilience of AI systems (ISO/IEC 24029 series), data quality and data governance (ISO/IEC 5259 series), functional safety (ISO/IEC TR 5469), and information security and cybersecurity (ISO/IEC 27001)[22].

In addition, several IEEE standards address the socio-technical dimensions of AI systems, with particular attention to explainability, responsibility, and the impact of AI-supported decision-

[19] Nastoska, A., Jancheska, B., Rizinski, M., and Trajanov, D., "Evaluating Trustworthiness in AI: Risks, Metrics, and Applications Across Industries," Electronics, 2025, doi:10.3390/electronics14132717

[20] Parlov, N., Mateša, B. and Mladinić, A., "Structuring AI Risk Management Framework: EU AI Act FRIA, GDPR DPIA and ISO 42001/23894," in Proceedings of the Mediterranean Conference on Embedded Computing (MECO), 2025.

[21] European Data Protection Supervisor (EDPS), Guidance for Risk Management of Artificial Intelligence Systems, 2025.

[22] International Organization for Standardization, ISO policy brief: Harnessing international standards for responsible AI development and governance, 2023.

making on individuals and society. For example, IEEE Std 7010-2020 provides methodologies for assessing the effects of autonomous and intelligent systems on human well-being[23], while IEEE P2976 focuses on requirements for explainable artificial intelligence[24]. IEEE P3396 addresses AI risk, safety, trustworthiness, and responsibility[25]. These IEEE standards further support the importance of an in-context risk assessment approach to the trustworthiness of AI systems that goes wider than just individual asset level measures.

At the European level, CEN and CENELEC, through their Joint Technical Committee JTC 21, are developing harmonised standards specifically supporting EU AI Act implementation for high-risk AI systems. Commission Implementing Decision C(2025) 3871 mandates the drafting of harmonised standards providing detailed technical specifications for the essential requirements on risk management (AI ACT, art. 9), data governance (AI ACT, art. 10), logging (AI ACT, art. 12), transparency (AI ACT, art. 13), human oversight (AI ACT, art. 14), accuracy, robustness and cybersecurity (AI ACT, art. 15), and quality management systems (AI ACT, art. 17), enabling compliant systems to benefit from presumption of conformity[26].

From ETSI, the Securing Artificial Intelligence (SAI) group develops AI related standards. Key references include ETSI TS 104 050 on AI threat ontology, ETSI GR SAI 009 on AI computing platform security, ETSI TS 104 224 on explicability and transparency of AI processing, and ETSI EN 304 223 establishing baseline cybersecurity requirements for AI models and systems.[27]

In addition, the recently renamed AI Safety Institute at NIST, now known as the Center for AI Standards and Innovation (CAISI)[28], is intended as an interface between government and industry for the testing, evaluation, and standardisation of commercial AI systems, with a strong emphasis on security and coordinated standards development. Its overarching aim is to cultivate a trustworthy AI ecosystem through voluntary, measurement-based standards that balance innovation with demonstrable risk reduction and it aligns with the NIST AI RMF Core Functions. For industry developers and evaluators, CAISI aims to provide collaboratively developed guidelines, benchmarks, and best practices that can be adopted voluntarily to demonstrate conformity with high-assurance security and safety expectations. It is also worth noting that for U.S. policy and national security experts it is intended to generate independent, unclassified

[23] Institute of Electrical and Electronics Engineers (IEEE), IEEE Std 7010-2020: Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being, 2020.
[24] Institute of Electrical and Electronics Engineers (IEEE), IEEE P2976: Standard for Explainable Artificial Intelligence (XAI), 2023.
[25] Institute of Electrical and Electronics Engineers (IEEE), IEEE P3396: Recommended Practice for Artificial Intelligence Risk, Safety, Trustworthiness, and Responsibility, 2023.
[26] European Commission, Commission Implementing Decision C(2025) 3871 on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation as regards high-risk AI systems in support of Regulation (EU) 2024/1689, 2025.
[27] ETSI, "Standards search: artificial intelligence," European Telecommunications Standards Institute, https://www.etsi.org/standards#page=1&search=artificial%20intelligence&title=1&etsiNumber=1&content=1&version=1&onApproval=1&published=1&withdrawn=1&historical=1&isCurrent=1&superseded=1&startDate=1988-01-15&endDate=2026-01-19&harmonized=0&keyword=&TB=&stdType=&frequency=&mandate=&collection=&sort=1
[28] https://www.nist.gov/caisi

evaluations of both domestic and foreign AI capabilities, including the identification of vulnerabilities, backdoors, and other covert, potentially malign behaviours in adversarial systems. In line with the recent political direction of travel in the U.S., it will be important for European organisations and institutions to remember that CAISI's purpose is to articulate technically grounded U.S. positions that "guard against burdensome and unnecessary regulation" while promoting U.S. leadership in global AI standards setting. Intended outcomes include consensus-based international standards that embed NIST's risk-based, measurement-centric approach and diffuse benefits of AI in ways aligned with U.S. economic and security interests, and as such the standards espoused may not fully align with EU policy or principles.

These suite of standards remain relevant for understanding the socio-technical environment in which AI systems are developed and assessed, but they are also complex, sometimes overlapping, and often challenging to implement as a whole.

2.4 Standards for Agentic AI systems

The recent trend towards Agentic AI systems - an autonomous AI that can plan, reason, and act in pursuit of goals with minimal human intervention - has raised the focus on trustworthiness, safety, interoperability, and governance. Agentic AI systems by definition are heavily connected to other AI and/or IT systems and present a new set of risks and threats, especially from cascading effects propagating across interlinked assets and systems with potentially significant impact resulting from an Agentic AI's ability to make decisions and initiate actions. In response to this there are emerging and evolving standardization efforts related to Agentic AIs.

It should be mentioned that there do not exist yet any generally accepted standards for Agentic AI governance and/or interoperability, but there is an emerging international ecosystem of (i) architectural proposals, (ii) open or vendor-backed agent protocols that act as proto-standards, and (iii) organisations like IEEE which work on Agentic AI interoperability. These may serve as a formal anchor for future Agentic-AI-specific standards and are examined below.

The Open Agentic Schema Framework[29] is a standardized schema system for defining and managing AI agent capabilities, interactions, and metadata. Launched in 2025, OASF provides a structured way to describe agent attributes, capabilities, and relationships using attribute-based taxonomies. OASF serves as the foundation for interoperable AI agent systems, enabling consistent definition and discovery of agent capabilities across distributed systems.

The Model Context Protocol[30] developed by Anthropic is a new standard for connecting AI assistants to the systems where data lives, including content repositories, business tools, and development environments. Its aim is to help frontier models produce better, more relevant responses. MCP aims to provide a universal, open standard for connecting AI systems with data sources, replacing fragmented integrations with a single protocol. The result is a simpler, more reliable way to give AI systems access to the data they need.

[29] https://github.com/agntcy/oasf?tab=readme-ov-file
[30] https://www.anthropic.com/news/model-context-protocol

AGNTCY[31] is an open-source collective building the critical infrastructure for AI agents to work together. It is maintained by companies like Cisco, LangChain and Calileo. It includes an open agent schema framework, an agent directory for discovering compatible agents and an agent connect protocol for cross-framework communication. ACP (Agent Connect Protocol)[32] as part of the AGNTCY initiative. It facilitates connections and interactions between AI agents, enabling more effective communication and collaboration.

The IEEE Computer Society is working on standards for agent communication and lifecycle management. Two specific ones are relevant: P3394 and P3428. The IEEE P3394[33] defines natural language interfaces that facilitate communication between Large Language Model (LLM) applications, agents, and human users. It defines a set of protocols and guidelines that enable applications and agents to effectively communicate with LLM enabled Agents. Thereby, the standard enables seamless interactions between multiple applications and agents. The IEEE P3428[34] is a standard for Large Language Model Agents for AI-powered Education and provides: the agentic AI building blocks and architectural elements that constitute an educational Large Language Model (LLM) agent; a communication protocol for interaction between different LLM agents and other components in the educational ecosystem; definitions of and standardised stages of deployment, operation, and retirement of educational LLM agents; guidelines for embedding and customizing foundation models within educational environments; and metrics and benchmarks for assessing the effectiveness, efficiency, and ethical considerations of LLM agents in educational settings.

In terms of guidelines the OWASP (Open Web Application Security Project), a global, non-profit foundation, published a report[35] that provides guidelines for securing autonomous systems, covering threats, vulnerabilities, and risk management. The report aims to serve as a practical guide for navigating the complexities of building, managing, and deploying agentic applications safely and effectively.

The Centre for AI Standards and Innovation (CAISI) 2025 also has a very recent dedicated workstream called the AI Agent Standards Initiative (AI ASI)[36]. This aims to ensure a trusted, interoperable and secure 'Agentic Frontier' through technical Agentic AI standards and open protocols. It intends to achieve these aims through a process of industry-led technical consultation and gap analysis, and through the development of open-source AI agent protocol ecosystems. However, it explicitly states that the primary purpose of this standardisation and protocol push is to "cement U.S. dominance at the technological frontier…[and]…maintain leadership on AI Agents at international standards bodies". As a result, CAISI and the AI ASI is therefore likely to prioritise U.S. conceptualisations of (un)necessary regulation, guardrailing and

[31] https://outshift.cisco.com/blog/building-the-internet-of-agents-introducing-the-agntcy
[32] https://spec.acp.agntcy.org
[33] https://standards.ieee.org/ieee/3394/11377/
[34] https://standards.ieee.org/ieee/3428/11489/
[35] https://genai.owasp.org/resource/state-of-agentic-ai-security-and-governance-1-0/
[36] https://www.nist.gov/caisi/ai-agent-standards-initiative

safeguarding, which may mean that the Agentic AI standards that do emerge from this body may not fit well with European needs and views.

Finally, the Agentic AI Foundation (AAIF)[37] was introduced in December 2025 by the Linux Foundation. AAIF includes contributions from leading technical projects including the previously mentioned Model Context Protocol (MCP) and provides a neutral, open foundation to ensure Agentic AI evolves transparently and collaboratively. Its key members include companies like Amazon Web Services, Anthropic, Google, Microsoft, OpenAI, Cisco, IBM, Salesforce, SAP, and Shopify.

These emerging and evolving Agentic AI standardisation efforts clearly indicate the level of interconnectedness present in such systems, further reinforcing the need for risk-based trustworthiness assessments to be conducted on a whole-system basis that can account for threats cascading through linked assets.

2.5 AI Risk Frameworks

The ENISA 'RISK MANAGEMENT STANDARDS: Analysis of Standardisation Requirements in Support of Cybersecurity Policy' report[38] provides a comprehensive analysis of international and European standards supporting risk management as a cornerstone of EU cybersecurity policy, particularly under the NIS Directive and Cybersecurity Act (Regulation (EU) 2019/881). The report compiles a curated list of applicable standards, conducts gap analyses, and identifies best practices, serving as key input for policy implementation in risk management domains.

Importantly, the gap analysis revealed that there is insufficient accounting for AI-specific threats. In response to this the FAITH AI TAF-v2 presents updates that explore LLM-specific threats (see 3.6.1) and Agentic AI-specific threats (see 3.6.4), and the FAITH Knowledgebase has been directly extended to include AI-specific threats (which will be reported in D2.4).

As an interesting extension to the notion of connected AI / IT systems and the need for socio-technical whole-system risk assessment, the ENISA Report: "Interoperable EU Risk Management Framework" (Updated December 2022)[39], proposes a methodology to assess and enhance the interoperability of risk management frameworks, addressing fragmentation in EU cybersecurity practices under the NIS Directive and Cybersecurity Act. In short, it explores ways to connect risk management methods. By evaluating prominent risk management methods against functional and non-functional features, it identifies pathways for coherent, combinable approaches to risk management. The report aims to provide a clear outcome regarding the potential for forming a coherent risk management framework that supports organisations across different sectors in the EU. In this context, *interoperability is defined as the ability of a risk management component or*

[37] https://aaif.io
[38] ENISA (2022). Risk Management Standards: Analysis of Standardisation Requirements in Support of Cybersecurity Policy. Luxembourg: Publications Office of the European Union.
[39] ENISA (2022). Interoperable EU Risk Management Framework. Athens: European Union Agency for Cybersecurity.

method to reuse information provided by another framework with the same ease, interfaces, and goals. The core innovation is a four-level interoperability scale (No/Low/Medium/High) applied to risk management lifecycle stages: context establishment, risk identification, assessment, treatment, monitoring, and review.

Accompanying the report is the ENISA Interoperable EU Risk Management Toolbox (2023)[40], which offers practical tools for operationalising risk management interoperability. It provides a knowledgebase and functional components and libraries. The knowledgebase provides the foundational data required to align disparate risk management efforts, including common terminologies, taxonomies and scales. The Functional Components and Libraries act as a bridge between the knowledgebase and RM methods. In response to these moves towards risk management interoperability, the FAITH AI TAF-v2 proposes the integration of the outputs of in-isolation risk assessments conducted at the asset level with the fine-tuning of in-context risk assessment conducted at the system level (see section 4.1).

2.6 Updates to AI Metrics

To ensure a holistic evaluation of AI trustworthiness, the FAITH AI TAF-v2 moves beyond simple accuracy scores and proposes various metrics on a trustworthiness oriented basis multi-layered metric framework. This section updates the state-of-the-art metrics, categorized by the specific risk dimension they addressTWA they are targetting. It enhances TAF-v1 by providing updated AI trustworthiness new metrics that enable the quantitative, consistent, technical and operational evaluation of trustworthiness risks across systems, lifecycle stages, and trustworthiness dimensions.

2.6.1 Metrics for the TWA: Fair with Harmful Bias Managed - Data Bias Detection

Trustworthy AI begins with the integrity of the input. Metrics in this category are designed to quantify statistical anomalies and representational disparities within the training corpora before they are encoded into model parameters. Rather than evaluating model predictions, these metrics analyse the intrinsic properties of the dataset itself. They generally operate by measuring the divergence between the observed distribution of sensitive attributes (such as age, gender, or ethnicity) and an ideal, balanced baseline. Fundamentally, they assess two forms of inequity: imbalance, where specific subgroups are quantitatively underrepresented, and correlation, where sensitive attributes improperly predict the target label (proxy bias).

The objective of this family of metrics is to detect historical bias, where the data accurately reflects a prejudiced past, and sampling bias, where the collection process failed to capture the diversity of the real world. By quantifying these disparities early, these metrics serve as a

[40] ENISA Interoperable EU Risk Management Toolbox

preventative control, ensuring that the model does not learn to generalize patterns based on systemic exclusion or noise.

2.6.2 Metrics for the TWA: Fair with Harmful Bias Managed - AI Model Bias Detection

While data metrics assess inputs, model bias metrics (often termed fairness metrics) evaluate the decision boundary created by the AI. These metrics determine whether the model treats different demographic groups equitably when making predictions. This family of metrics generally operates by comparing error rates and prediction outcomes across protected groups. They test for group fairness (e.g. does the model accept loans for men and women at the same rate?) and individual fairness (e.g. do similar individuals receive similar predictions regardless of sensitive attributes?). A key conceptual distinction here is the trade-off between equality of opportunity (ensuring true positives are equal) and equality of outcome (ensuring acceptance rates are equal).

These metrics are critical for identifying algorithmic amplification, where a model not only learns the biases present in the training data but exaggerates them to minimise a loss function. By monitoring these values during validation, developers can identify if the model is optimising for accuracy at the expense of equity for specific sub-populations.

Table 1: AI Model Bias Metrics

No	Metric (acronym)	Short description	Reference
1	Difference in Positive Proportions in Predicted Labels (DPPL)	Measures the difference in the probability of a positive prediction between the favoured facet and disfavoured facet	https://patents.google.com/patent/US20220172004A1/en
2	Disparate Impact (DI)	Computes the ratio of positive prediction rates between the disfavoured and favoured facets to detect adverse impact	https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-post-training-bias-metric-di.html
3	Conditional Demographic Disparity in Predicted Labels (CDDPL)	The weighted sum of disparities calculated within specific subgroups to control for confounding variables	https://pages.awscloud.com/rs/112-TZM-766/images/Fairness.Measures.for.Machine.Learning.in.Finance.pdf
4	Counterfactual Flip Test (FT)	Calculates the rate at which the model's prediction changes when only the sensitive attribute in input x is flipped to create x'	https://par.nsf.gov/servlets/purl/10165049
5	Accuracy Difference (AD)	The difference between the prediction accuracy for the favoured facet and the disfavoured facet	https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-post-training-bias-metric-ad.html
6	Recall Difference (RD)	Measures the difference in the True Positive Rate (Recall) between the favoured and disfavoured facets (Equality of Opportunity)	https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-post-training-bias-metric-rd.html
7	Difference in Conditional Acceptance (DCAcc)	Measures the difference in acceptance rates between the favoured and disfavoured facets conditional on the true positive class	https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-post-training-bias-metric-dcacc.html

8	Difference in Acceptance Rates (DAR)	Measures the difference in observed positive rates between facets in the training data (Pre-training bias)	https://pages.awscloud.com/rs/112-TZM-766/images/Amazon.AI.Fairness.and.Explainability.Whitepaper.pdf
9	Specificity Difference (SD)	Compares the True Negative Rate (Specificity) between facets to check if negatives are correctly identified at equal rates	https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-post-training-bias-metric-sd.html
10	Difference in Conditional Rejection (DCR)	Measures the difference in rejection rates between the favoured and disfavoured facets conditional on the true negative class	https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-post-training-bias-metric-dcr.html
11	Difference in Rejection Rates (DRR)	Measures the difference in the rate of negative predictions between the favoured and disfavoured facets	https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-post-training-bias-metric-drr.html
12	Treatment Equality (TE)	Measures the difference in the ratio of False Negatives to False Positives between facets, balancing the types of errors made	https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-post-training-bias-metric-te.html
13	Generalized Entropy Index (GE)	Measures inequality in the distribution of benefit across individuals	https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-post-training-bias-metric-ge.html

2.6.3 Metrics for the TWA: Privacy-Enhanced - Data Privacy Assessment

In the context of trustworthy AI, privacy metrics serve to quantify the risk of sensitive information leakage through the model's outputs or parameters. Unlike security controls that restrict access, these metrics evaluate the information theoretic boundaries of what the model has memorized.

The core concept behind this family of metrics is ‘distinguishability’. They measure how easy it is for an adversary to determine whether a specific individual’s data was used to train the model (membership inference) or to reconstruct the original attributes of a training sample (model inversion). This category also encompasses metrics derived from Differential Privacy, which quantify the maximum probability that a model's output will change if a single individual is added or removed from the dataset.

By calculating these values, organizations can assess the trade-off between model utility and data confidentiality. High scores in this category indicate that the model has over-memorized the training data rather than learning generalizable patterns, posing a direct violation of privacy regulations like the GDPR.

Table 2: Data Privacy Metrics

No	Metric (acronym)	Short description	Reference
1	Membership Inference Attack Success Rate (MI Asr)	Calculates the accuracy of an adversary in correctly predicting the	https://ieeexplore.ieee.org/document/7958568

		membership status ($m_i \in \{0,1\}$) of input x_i	
2	Attribute Inference / Model Inversion Success	Measures the probability that an adversary can correctly reconstruct sensitive attribute given the model output and non-sensitive features	https://dl.acm.org/doi/10.1145/2810103.2813677
3	Differential Privacy Budget (ε, δ)	The bound parameters ensuring the output distribution on dataset D is statistically indistinguishable (within e^{ϵ},δ) from neighbor D'.	https://www.emerald.com/fttcs/article-abstract/9/3-4/211/1332491/The-Algorithmic-Foundations-of-Differential
4	Overconfidence / Calibration	Quantifies the divergence in average prediction confidence between training members and non-members; a high gap signals memorization.	https://www.usenix.org/system/files/sec21-carlini-extracting.pdf

2.6.4 Metrics for the TWA: Explainable and Interpretable - Explainability Evaluation

While performance metrics tell us if a model is correct, explainability metrics evaluate the validity and quality of the reasoning behind those predictions. These metrics assess the explanations (such as feature attribution maps) generated by auxiliary tools to ensure they are not misleading.

This family of metrics focuses on three conceptual pillars: Fidelity, Stability, and Complexity. Fidelity metrics measure whether the explanation truthfully reflects the model's decision process (e.g., if the explanation says a feature is important, removing it should hurt accuracy). Stability metrics assess consistency, checking if similar inputs yield similar explanations, or if the reasoning is brittle to noise. Complexity metrics evaluate the cognitive load of the explanation, operating on the assumption that concise, sparse explanations are more interpretable to humans than dense, high-dimensional ones.

The role of these metrics is to prevent plausible but wrong explanations, cases where an AI system offers a convincing justification that hides the true, potentially biased, logic it actually used.

Table 3: Explainability Metrics

No	Metric (acronym)	Short description	Reference
1	Faithfulness (Fidelity)	Measures correlation between feature importance and the drop in model output when that feature is masked.	https://jmlr.org/papers/v24/22-0142.html
2	Stability (Local Lipschitz)	Evaluates the ratio of the change in explanation E(x) to the change in input δ.	https://arxiv.org/abs/1806.08049

3	Complexity (Sparsity)	Quantifies cognitive load by counting the number of non-zero features (L_0 norm) returned in the explanation vector E(x).	https://arxiv.org/abs/1602.04938
4	Monotonicity	Verifies that adding features with positive attribution strictly increases or maintains the model's prediction score.	https://arxiv.org/abs/1703.01365
5	Identity	A deterministic check ensuring that if two inputs are identical, their generated explanations are strictly equal.	https://arxiv.org/abs/1810.03292
6	Max-Sensitivity	Measures the maximum deviation in the explanation given an input perturbation within a radius	https://arxiv.org/abs/2412.10942
7	SHAP (Shapley Additive Explanations)	Feature importance based on game theory (Shapley values)	https://dl.acm.org/doi/10.5555/3295222.3295230
8	LIME (Local Interpretable Model-agnostic Explanations)	Local surrogate model accuracy and fidelity	https://arxiv.org/abs/1602.04938
9	Global Feature Importance Spread (GFIS)	Measures the standard deviation of the mean importance scores across all features	https://arxiv.org/pdf/2302.12094

2.6.5 Metrics for the TWA: Secure and Resilient - Adversarial Robustness and Security

Trustworthy AI must be resilient to active malice. Metrics in this category quantify a model's ability to maintain correct performance in the presence of adversarial attacks, inputs intentionally modified to cause errors. The operational logic of these metrics centers on perturbation thresholds. They typically measure the distance (or magnitude of noise) required to flip a model's prediction. A higher distance implies a more robust model. This family also includes attack success rates which empirically test the model against known attack libraries, and certified radius metrics, which provide mathematical guarantees of stability around a given input point.

Unlike standard accuracy, which assumes benign distribution of data, these metrics evaluate the model's worst-case behavior. They are essential for safety-critical systems, determining whether a model can be easily fooled by imperceptible noise or specific trigger patterns introduced by an attacker.

Table 4: Adversarial Robustness and Security Metrics

No	Metric (acronym)	Short description	Reference
1	Attack Success Rate (ASR)	Calculates the ratio of successful adversarial attacks to the total number of attempts	https://link.springer.com/article/10.1007/s10462-025-11147-4
2	Minimal Perturbation Distance (L_p)	Measures the minimum distortion (using Euclidean or Max norm) needed to flip a prediction	https://arxiv.org/abs/1511.04599
3	CLEVER Score	Estimates a theoretical lower bound for robustness (minimum perturbation) using the local Lipschitz constant	https://arxiv.org/abs/1801.10578
4	Empirical Robustness (ER)	Computes the mean of the minimal perturbations required to fool the model, often normalized by input magnitude	https://link.springer.com/article/10.1007/s10462-025-11147-4
5	Certified Radius	Determines the radius R around an input where the prediction is guaranteed to be stable (using Randomized Smoothing)	https://arxiv.org/abs/1902.02918
6	Loss Sensitivity	Assesses how sensitive the model's loss function L is to small variations in the input data (Frobenius norm of the gradient)	https://link.springer.com/article/10.1007/s10462-025-11147-4

2.6.6 Metrics for Efficiency and Sustainability (Green AI)

Metrics for efficiency and sustainability address the computational, energy, and environmental footprint of AI systems across their lifecycle. While these metrics are not directly mapped to a single Trustworthiness Attribute (TWA), they are closely related to the broader notion of Safe AI, particularly when safety is interpreted in a systemic and long-term sense, including societal and environmental impact. From a trustworthiness perspective, inefficient AI systems may introduce indirect risks, such as excessive resource consumption, limited deployability in constrained environments, or disproportionate environmental costs relative to their benefit. These factors can affect the overall acceptability, scalability, and responsible use of AI systems, especially in large-scale or continuous deployment settings. Therefore, efficiency and sustainability metrics complement the core trustworthiness dimensions by enabling the assessment of resource-aware and environmentally responsible AI, supporting a more holistic, lifecycle-oriented view of AI safety and risk management.

Trustworthiness encompasses not only the accuracy of a system but also its societal footprint and operational viability. This family of metrics evaluates the computational cost, environmental impact, and resource efficiency of AI models throughout their lifecycle. The core concept here is

the 'accuracy-efficiency trade-off'. These metrics quantify the resources consumed, such as energy (kWh), carbon emissions (CO2e), and compute operations (FLOPs), relative to the performance gain. This category also includes operational metrics like inference latency and throughput, which determine if a model is fast enough for real-time safety systems.

By monitoring these values, organizations move towards Green AI, prioritising models that achieve sustainable performance over those that burn excessive energy for marginal accuracy gains. This is critical for assessing the long-term deployment risks and the broader ethical responsibility of the AI system regarding climate impact.

Table 5: Efficiency and Sustainability Metrics

No	Metric (acronym)	Short description	Reference
1	Power Usage Effectiveness (PUE)	Ratio of total facility energy to energy reaching the computing equipment	https://datacenters.lbl.gov/sites/default/files/WP49-PUE%20A%20Comprehensive%20Examination%20of%20the%20Metric_v6.pdf
2	Carbon Equivalent Emissions (CO_2eq)	Estimates CO_2 equivalent emissions from CPU, GPU, and RAM energy consumption weighted by grid emission intensity and data-center efficiency	https://link.springer.com/article/10.1134/S1064562422060230
3	Energy Consumption (kWh)	Measures total energy drawn by integrating power usage over the duration of operation	https://arxiv.org/abs/1907.10597
4	Floating Point Operations (FLOPs)	Counts the total arithmetic operations (additions/multiplications) across all model layers	https://arxiv.org/abs/2107.11949
5	Inference Latency	Measures the real-time duration between the input request and the model's generated prediction	https://arxiv.org/pdf/2407.07000
6	Model Parameters (Size)	Counts the number of trainable weights in the model	https://arxiv.org/abs/2001.08361
7	Green AI (Efficiency)	Framework treating efficiency as an optimization term alongside accuracy in the loss function	https://arxiv.org/abs/1907.10597

2.6.7 Metrics for Uncertainty and Calibration

A trustworthy AI system must be aware of its own limitations. Metrics in this category assess calibration, the alignment between the model's confidence scores and its actual accuracy. Fundamentally, these metrics answer the question: "When the model says it is 90% sure, is it actually correct 90% of the time?". They measure the discrepancy between predicted probabilities and empirical frequency (e.g., Expected Calibration Error). This family also includes

uncertainty quantification metrics, such as the width of prediction intervals in regression tasks or the detection rate of out-of-distribution (OOD) data.

These metrics are the fail-safes of AI deployment. They ensure that high-stakes decisions are not made based on overconfident predictions. A well-calibrated model will output low confidence when facing unfamiliar data, allowing the system to trigger a human-in-the-loop hand-off rather than failing silently.

Table 6: Calibration Metrics

No	Metric (acronym)	Short description	Reference
1	Expected Calibration Error (ECE)	The average difference between predicted confidence and actual accuracy	https://proceedings.mlr.press/v70/guo17a.html
2	Prediction Interval Coverage Probability (PICP)	The percentage of ground truth values that fall within the predicted upper and lower bounds	https://ieeexplore.ieee.org/document/5966350
3	Brier Score	The mean squared difference between predicted probability and actual outcome	https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml
4	OOD Detection Rate (AUROC)	The Area Under the ROC curve for the binary classification task of distinguishing in-distribution vs. out-of-distribution data	https://arxiv.org/abs/1610.02136

3 An In-context Risk Assessment-Based Approach to AI Trustworthiness

This Chapter extends the theoretical and methodological foundation of TAF-v1 such that TAF-v2 can account for socio-technical in-context system-level trustworthiness assessment. It introduces key concepts such as interconnected systems, impact levels as a way to customise context, the importance of humans in the system, cascading threats, nuanced AI Model classifications and their associated risks, and trustworthiness trade-offs, which directly underpin the enhancements introduced in the framework.

3.1 Interconnected and Interdependent Systems

In response to the increasing focus of regulatory and standards requirements on socio-technical whole-system risk assessment, a critical extension from the FAITH AI TAF-v1 is the introduction throughout TAF-v2 of the concept and implications of risk assessing AI assets within the context in which they are designed, developed and deployed. An in-isolation risk assessment aims to identify intrinsic, asset-internal risk properties, independent of system integration and socio-technical dependencies. In this context, model and data assets represent the core functional determinants of AI system behaviour. They define the statistical and computational behaviour of AI systems. Data shapes learned representations, biases, and generalisation properties, while models encode decision logic, representations, and generative capabilities. Consequently, many fundamental AI-related threats, including for example bias, robustness failures, overfitting, hallucinations, or model misuse potential, can be conceptually traced to properties intrinsic to model architectures and data.

However, unlike an in-isolation risk assessment, an in-context risk assessment explicitly addresses systemic and cascading risks, which emerge from interactions among multiple assets, lifecycle stages, and socio-technical components. This represents a shift from an in-isolation asset-level risk assessment to an in-context whole-system risk assessment, where AI assets are conceptualized as interconnected and interdependent on people, other AI and software processes, data, hardware, physical locations and jurisdictions, and regulatory requirements. As ENISA's Multilayer Framework[41] states, the *"main challenges to security […] imply a dynamic analysis of the threats, both along the life cycle and in the interrelations within other blocks of an ICT infrastructure"*. Equally, the NIST AI RMF states that *"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."*

Risk in interconnected and interdependent systems (e.g. a system-under-test in a specific context) is fundamentally different from risk in a standalone model or application (e.g. an asset

[41] ENISA European Union Agency for Cybersecurity. 2023, p.24. *Multilayer Framework for Good Cybersecurity Practices for AI* available on https://www.enisa.europa.eu/sites/default/files/publications/Multilayer%20Framework%20for%20Good%20Cybersecurity%20Practices%20for%20AI.pdf

in isolation) because hazards propagate through couplings, shared resources, and feedback loops[42]. In such contexts, the overall risk profile results from the interactions among components, organisations, and processes. Consequently, outcomes may be non-linear and emergent and cannot be reliably inferred by evaluating each element independently. A small disturbance can escalate when dependencies transmit disruption across technical, human, and organisational boundaries, for instance when a common third-party service failure simultaneously undermines operations, monitoring, and recovery capabilities across multiple teams[43]. This systemic view is particularly relevant for AI-enabled systems, where functional boundaries routinely span technical components, organisational roles, and operational processes.

In AI-enabled contexts, interaction-driven risk is particularly significant because the AI system typically comprises an assemblage of interconnected components and stakeholders, including data generation and curation, model development, deployment infrastructure, and human decision-making regarding its outputs. Consequently, harms often emerge when minor mismatches or degradations, such as those in data quality, operational assumptions, or organizational handoffs, propagate throughout the system lifecycle and escalate into downstream failures that are challenging to diagnose from the perspective of any single component[44]. Empirical studies of high-stakes AI refer to this phenomenon as data cascades, in which early data and process issues accumulate and expand through interconnected workflows, ultimately resulting in systemic failures during deployment. Additionally, research indicates that modern machine learning pipelines are frequently underspecified, so that many configurations may appear equally effective during validation but behave unpredictably after deployment. This further increases the system's sensitivity to interactions among components rather than to isolated model performance[45] [46].

Consequently, these characteristics show that managing risk in interconnected and interdependent AI-enabled systems requires a whole-of-system approach that treats boundaries, interfaces, and operational context as risk causes. In practice, this changes emphasis from one-time, component-level assurance where assets are treated in isolation from their connections to

[42] Min, B.H. and Borch, C., 2022. Systemic failures and organizational risk management in algorithmic trading: Normal accidents and high reliability in financial markets. Social studies of science, 52(2), pp.277-302

[43] Duan, D., Lv, C., Si, S., Wang, Z., Li, D., Gao, J., Havlin, S., Stanley, H.E. and Boccaletti, S., 2019. Universal behavior of cascading failures in interdependent networks. Proceedings of the National Academy of Sciences, 116(45), pp.22452-22457.

[44] Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. and Aroyo, L.M., 2021, May. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-15).

[45] AI, N., 2023. Artificial intelligence risk management framework (AI RMF 1.0). URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai, pp.100-1.

[46] D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M.D. and Hormozdiari, F., 2022. Underspecification presents challenges for credibility in modern machine learning. Journal of Machine Learning Research, 23(226), pp.1-61.

other systems, toward lifecycle governance that recognises dependencies, identifies risks to the whole system arising from cascading effects, and provides controls and control strategies that are targeted to have the maximum mitigation effect through the minimisation of cascade effects.

3.2 Importance of Impact in In-context Risk Assessment

The first key feature of an in-context approach to AI risk and trustworthiness assessment involves the consideration of the wider socio-technical system in which an AI system is located. Two keys differences with this when compared with risk assessment of the AI system on its own are that consideration is given to incoming threats from the wider socio-technical system that can affect the AI system and that consideration is given to consequences and impacts of risks associated with the AI system propagating to and affecting other components and, importantly, people within the wider socio-technical system. Therefore, for in-context risk assessment, impact is important.

Following this, understanding impact is central to in-context risk assessment. It provides the basis for evaluating the consequences of risks of the AI system and its behaviour within the socio-technical environment, both positive and negative. The NIST AI RMF defines risk as a composite of likelihood and the magnitude of consequences, emphasizing that impacts may affect individuals, groups, organisations, society, and even the broader ecosystem. Among many others, it includes harms to civil liberties, rights, and safety, and suggests that assessing these impacts is crucial for building trustworthy AI systems. With impacts included within risk calculations, stakeholders and concerned individuals gain a better and more comprehensive understanding of how system failures, misclassifications, or biases may manifest in real-world contexts and how these can impact affected parties.

The NIST AI RMF also highlights that AI risks are dynamic and therefore that human users might overestimate AI reliability or objectivity, making it even more important to contextualise impacts during development and deployment so that their importance and effects are accurately captured during in-context risk assessments. Pre-empting these overestimations and considering failures, biases and harms helps strengthen AI systems, the socio-technical systems they sit within, as well as compliance and mitigation strategies.

Similar to the NIST AI RMF, the ENISA FAICP[47] gives impact as a core element of AI cybersecurity risk assessment. It too acknowledges the wider socio-technical nature in which AI systems are placed, and the cascading threats, risks and consequences that arise from misuse, vulnerabilities and inadequate controls. The layered approach of the FAICP considers impacts across the lifecycle stages of an AI system in context. This then allows for impacts to be considered in-context and, critically, considers how they change throughout the different lifecycle stages.

[47] ENISA European Union Agency for Cybersecurity. 2023, p.24. *Multilayer Framework for Good Cybersecurity Practices for AI* available on https://www.enisa.europa.eu/sites/default/files/publications/Multilayer%20Framework%20for%20Good%20Cybersecurity%20Practices%20for%20AI.pdf

With impact playing such an important role in different facets of in-context risk assessment, regular in-context risk assessment, involving domain experts alongside risk assessors and developer teams, is needed to sufficiently monitor and assess it, in order that the correct severity and scope of different potential impacts resulting from the same or newly identified risks are accounted for. This then ensures that the in-context risk assessment captures different and nuanced domain- and sector-relevant consequences and impacts, as well as enabling the recommendation of context-appropriate mitigations and Control measures. Meaningful in-context risk assessment of AI systems depends on a deep understanding of how impacts unfold within specific contexts of use, an understanding requiring domain and organisational expertise and appropriate risk assessment tools.

3.3 Humans in the System

The second key feature of an in-context approach to AI risk and trustworthiness assessment involves the consideration of the humans who interact with the technologies. In complex and connected AI systems, human actors play multiple roles across different stages of the AI lifecycle, and their actions can significantly influence trustworthiness outcomes. As introduced in D2.1, humans are involved in the design, development, deployment, operation, and governance of AI systems in interconnected socio-technical environments.

Different human roles, such as developers, data engineers, system operators, decision-makers, and end-users, interact with AI assets in distinct ways and are exposed to different types of threats and vulnerabilities. These human-related risks may arise from both unintentional behaviour (e.g. errors, misconfigurations, cognitive biases) and intentional exploitation (e.g. social engineering or phishing attacks targeting privileged users).

In connected systems, human-related threats do not remain localised. For example, a successful phishing attack against a system administrator may lead to credential compromise, enabling unauthorised access to multiple interconnected AI components. This can result in cascading effects, such as manipulation of training data, alteration of model parameters, or disabling of monitoring mechanisms, thereby amplifying the impact across the system. Similarly, errors introduced during development or deployment, such as incorrect configuration of access controls or inadequate validation of data sources, may propagate through downstream components and affect system behaviour in production.

Human vulnerabilities also vary across lifecycle stages. During development, risks may relate to insufficient awareness of bias, security, or robustness requirements. During deployment and operation, over-reliance on AI outputs, inadequate oversight, or insufficient incident response capabilities may increase the likelihood and impact of failures. At governance and oversight levels, unclear accountability or weak escalation procedures may delay corrective actions, further exacerbating cascading risks.

Assessing human-related risks therefore requires a lifecycle-aware and role-sensitive approach that explicitly considers how threats targeting humans, such as phishing, insider misuse, or social engineering, can exploit organisational and technical interdependencies. By integrating these considerations into the trustworthiness assessment, FAITH AI-TAF v2 ensures that human-related

threats and vulnerabilities are systematically identified and evaluated alongside technical risks, supporting a more realistic and comprehensive understanding of trustworthiness in connected AI systems.

3.4 Cascading Threats

The third, and perhaps most critical, key feature of an in-context approach to AI risk and trustworthiness assessment involves the consideration of the opportunity for threats to one asset to propagate to connected assets either directly or through changed behaviours in the primary asset. These cascading threats denote an important aspect of AI trustworthiness that goes beyond the vulnerability of a single system to the interrelations of contemporary AI architectures. Complex AI ecosystems, especially Agentic AI systems for example, are not often single systems that just exist, but instead they are components within a larger network where data moves between systems, models exchange parameters, and decisions made in one system affect other systems on which they depend or through which they act[48] [49].

A cascading threat happens when a security threat, adversarial attack, or system failure in one assets, known as the Primary Threat, is transmitted through system dependencies to downstream assets or other connected systems, resulting in a chain of Secondary Threats propagating across interconnected assets, systems, or decision-making processes. Cascading threats are especially applicable in situations where AI systems become part of critical infrastructure. Once an attacker manages to break one of the points of such a network, be it due to adversarial manipulation, data poisoning, or model extraction, the resulting consequences are not confined there. Rather they trickle down the system, impacting on the dependent models, decision-making, as well as finally the end-users or the critical applications[50] [51]. As an example, within autonomous vehicle networks, where the robustness of a perception system has been undermined by adversarial manipulation of sensor input, the cascading effect may negatively affect decision-making by other, dependent systems, causing safety-critical failures throughout the network of vehicles connected to each other[52].

Various structural properties of modern AI systems give rise to cascading threats: (1) data dependencies, where data outputs are used as inputs to other systems[53]; (2) parameter sharing, in which systems heavily rely on federated learning, and poisoned local models are applied to the

[48] J. Lin and L. Zhang, “From Failure Modes to Reliability Awareness in Generative and Agentic AI System,” arXiv.org, Oct. 2025, doi: 10.48550/arXiv.2511.05511.

[49] A. Reid, S. O’’Callaghan, L. Carroll, and T. Caetano, “Risk Analysis Techniques for Governed LLM-based Multi-Agent Systems,” arXiv.org, Aug. 2025, doi: 10.48550/arXiv.2508.05687.

[50] L. D. Valdez et al., “Cascading failures in complex networks,” J. Complex Networks, Apr. 2020, doi: 10.1093/comnet/cnaa013.

[51] J. Zhao, D. Li, H. Sanhedrai, R. Cohen, and S. Havlin, “Spatio-temporal propagation of cascading overload failures in spatially embedded networks,” Nature Communications, Jan. 2016, doi: 10.1038/ncomms10094.

[52] E. Nowroozi, I. Haider, R. Taheri, and M. Conti, “Federated Learning Under Attack: Exposing Vulnerabilities Through Data Poisoning Attacks in Computer Networks,” IEEE Transactions on Network and Service Management, Mar. 2024, doi: 10.1109/TNSM.2025.3525554.

[53] X. Wu et al., “Hybrid learning-based fault prediction and cascading failure mitigation in multi-network energy systems,” Scientific Reports, Sep. 2025, doi: 10.1038/s41598-025-10304-7.

global model[54]; (3) decision dependencies, in which decision outputs of one AI system are used as input to other systems[55]; and (4) shared vulnerabilities, where compromised shared resources are used by all systems that are dependent on it[56].

At a high-level, it is possible to map cascading threats to the seven FAITH Trustworthiness Attributes (TWAs) defined in NIST AI RMF and detailed in TAF-v1, as follows:

Safe

The design and monitoring of AI systems should, in addition to operating safely in isolation, be able to work safely in integrated environments where failure or unsafe outputs can spread to other systems depending on it, with increasing risk of harm to people, societies and the environment. Safety mechanisms should sufficiently consider the potential for the multiplication of unsafe situations across assets and system boundaries so that safety violations leading to physical, mental, rights violations, societal, and environmental harms do not propagate to generate or compound those harms further downstream or in wider systems[57]. In this regard safe defaults, regular mandatory human-in-the-loop intervention points, and complementary risk assessments (such as environmental or economic impact assessment) can all be beneficial.[58].

Secure & Resilient

Security and resiliency should be implemented throughout the whole chain of interdependent systems against the spread of attacks[59] through whole-system cybersecurity risk assessment that evaluates the security position of the complete supply chain and interdependent system network, not just the independent components[60]. AI systems should include internal security boundaries to ensure that attackers do not use weaknesses in upstream systems to compromise the downstream systems. The mechanism of resilience should take into consideration correlated failures, wherein attacks on two or more systems happen either in parallel, or in series[61]. This can

[54] A. Yazdinejad, A. Dehghantanha, H. Karimipour, G. Srivastava, and R. Parizi, "A Robust Privacy-Preserving Federated Learning Model Against Model Poisoning Attacks," IEEE Transactions on Information Forensics and Security, 2024, doi: 10.1109/TIFS.2024.3420126.

[55] M. Li, C. Tse, D. Liu, and X. Zhang, "Cascading Failure Propagation and Mitigation Strategies in Power Systems," IEEE Systems Journal, Jun. 2023, doi: 10.1109/JSYST.2023.3248044.

[56] A. Salehpour, I. Al-Anbagi, K. Yow, and X. Cheng, "Modeling Cascading Failures in Coupled Smart Grid Networks," IEEE Access, 2022, doi: 10.1109/access.2022.3194989.

[57] B. Venkatasubramanian, C. Laoudias, and M. Panteli, "Cascading Failures and Resilience in Interdependent Critical Infrastructures: A Dynamic Approach," IEEE Systems Journal, Dec. 2025, doi: 10.1109/JSYST.2025.3629327.

[58] Z. Guo, X. Su, K. Sun, B. Park, and S. Simunovic, "Analysis and Mitigation of Cascading Failures Using a Stochastic Interaction Graph with Eigen-analysis," IEEE Power & Energy Society General Meeting, Jul. 2025, doi: 10.1109/PESGM52009.2025.11225773.

[59] I. Adabara, B. O. Sadiq, A. N. Shuaibu, Y. I. Danjuma, and V. Maninti, "Trustworthy agentic AI systems: a cross-layer review of architectures, threat models, and governance strategies for real-world deployment," F1000Research, Sep. 2025, doi: 10.12688/f1000research.169927.1.

[60] M. Giacalone, "Quantifying Propagation Risk in Distributed Critical Infrastructures: A Unified Framework for AI Failures and GPS Spoofing," IEEE Access, 2026, doi: 10.1109/ACCESS.2026.3655246.

[61] D. Zhao, Z. Wang, G. Xiao, B. Gao, and L. Wang, "The robustness of interdependent networks under the interplay between cascading failures and virus propagation," arXiv.org, Aug. 2016, doi: 10.1209/0295-5075/115/58004.

involve installing a range of safety barriers such as faut tolerance defaults and error detection; redundancy; circuit breakers and mechanisms for rollbacks that are able to contain safety problems before they spread across the whole-system; resilience strategies which identify when attacks spread across system dependencies; and compartmentalisation to restrict the growth of the attack surface[62].

Explainable & Interpretable

Explainability and interpretability when applied to interlinked assets can be interpreted as the impact of cascading decisions from one system to related systems[63] where not only can erroneous decisions negatively impact those linked systems, but where explainability paths are also obscured and obfuscated. When the decision outputs of one interpretable system are the inputs of downstream systems, the flow of reasoning should be clear and traceable to all the components. This involves recording the cascading of decisions through system networks, preserving the fidelity of explanation in the manner predictions cascade through multiple steps, and giving end-users the ability to view not only the outputs of specific models but also the entire decision path through to the inputs used to generate those outputs. In this way it becomes possible to identify where an upstream erroneous decision output has resulted in an amplified downstream error. Further, errors in upstream models or adversarial manipulations should be explained in a way that should clarify how the validity of downstream interpretability could be undermined, and the system should give warning signs to indicate that cascading failures may have caused explanations to be invalid[64].

Privacy-enhanced

Privacy protection should be enhanced beyond asset and even system boundaries to ensure privacy is not threatened by a privacy violation in one system impacting the privacy in other systems which are dependent[65]. Especially in federated learning or collaborative learning contexts (where devices train AI models locally and where federated learning is a privacy-enhancing strategy) privacy protocols should ensure that if the local model is poisoned the corrupted model can't leak or send private information to the main "global" or shared AI system[66]. Privacy-aware AI systems should have mechanisms that identify when privacy violations in the upstream systems make the downstream systems more vulnerable, especially in

[62] T. Liu and F. Liu, “Graph Neural Networks for Evaluating the Reliability and Resilience of Infrastructure Systems: A Systematic Review of Models, Applications, and Future Directions,” IEEE Access, 2025, doi: 10.1109/ACCESS.2025.3611333.

[63] S. Raza, R. Sapkota, M. Karkee, and C. Emmanouilidis, “Responsible Agentic Reasoning and AI Agents: A Critical Survey,” Robotics, Oct. 2025, doi: 10.70777/si.v2i6.16169.

[64] R. K. Vijayaraj, “Embeddings to Diagnosis: Latent Fragility under Agentic Perturbations in Clinical LLMs,” arXiv.org, Jul. 2025, doi: 10.48550/arXiv.2507.21188.

[65] J. Wu, F. Luo, T. Sun, H. Wang, and W. Zhang, “Privacy-Preserving Federated Learning Scheme With Mitigating Model Poisoning Attacks: Vulnerabilities and Countermeasures,” IEEE Transactions on Dependable and Secure Computing, Jun. 2025, doi: 10.1109/TDSC.2025.3617070.

[66] H. Holla, A. A. Sasikumar, C. Gutti, K. Thumula, and H. Gogineni, “Advanced Privacy and Security Techniques in Federated Learning Against Sophisticated Attacks,” International Symposium on Digital Forensics and Security, Apr. 2025, doi: 10.1109/ISDFS65363.2025.11012115.

a situation where data is traversed by several processing stages or when sharing model parameters. The guarantees of differential privacy should be ensured in cascading computation and cumulative loss of privacy should be noticed with the flow of data through connected systems[67]. Also, systems have to take into consideration how privacy risks are magnified when anonymized information across several sources is recombined downstream.

Fair

Fairness guarantees should be applied to cascading decisions and data where the results and outputs of a single model affect the results of the dependent model and/or the wider system as a whole[68]. Biases of upstream systems may be increased or even reformulated as they run the downstream elements and may generate more or more serious discriminative patterns. Equitable AI systems should have in place mechanisms to ensure that bias amplification of chains of systems is detected, correlated fairness violations, and fairness constraints are enforced through the entire decision pipeline. Organizations should evaluate the implication of fairness on the whole system architecture, considering how bias and/or a lack of equality and diversity data and outputs of primary models can be multiplied in subsequent decisions and outputs to cause compounded bias in the ultimate decision or output. Constant observation and assessment of fairness should not be isolated to the components but should encompass the whole network of interdependent systems[69].

Accountable & Transparent

Accountability and transparency should be spread out over the whole network of dependent and interconnected systems[70]. The auditing processes should be able to trace the decision chains between systems and record the means in which decisions are passed through interconnected systems and at what points cascading failures may start. Traceability requirements should include data lineage, model lineage, and decision path tracking throughout all system layers, which allow the auditor to know which upstream components caused the failures of downstream systems. Maps of system dependencies need to be transparently documented, the propagation mechanisms by which failures are propagated clearly defined and understood, and accountability structures should be put in place that define responsibility when failures begin in one system and are reflected as failures in dependent systems[71]. The audit done by external and internal auditors

[67] X. Li, M. Wen, S. He, R. Lu, and L. Wang, "A Privacy-Preserving Federated Learning Scheme Against Poisoning Attacks in Smart Grid," IEEE Internet of Things Journal, May 2024, doi: 10.1109/JIOT.2024.3365142.
[68] S. I. A. Meerza and J. Liu, "EAB-FL: Exacerbating Algorithmic Bias through Model Poisoning Attacks in Federated Learning," International Joint Conference on Artificial Intelligence, Aug. 2024, doi: 10.24963/ijcai.2024/51.
[69] C. R. Kumar, S. Shanmuk, P. N. Srinivas, S. V. Lankalapalli, and D. Dwibedy, "A Data-Driven Probabilistic Framework for Cascading Urban Risk Analysis Using Bayesian Networks," arXiv.org, May 2025, doi: 10.48550/arXiv.2505.06281.
[70] J. Li, F. Bnaben, J. Gou, and W. Mu, "Modeling cascading effects in collaborative systems: a formal risk interdependency framework," Journal of Modelling in Management, Feb. 2025, doi: 10.1108/jm2-09-2024-0313.
[71] K. Singh, V. K. Chandrasekar, W. Zou, J. Kurths, and D. V. Senthilkumar, "Graph coloring framework to mitigate cascading failure in complex networks," Communications Physics, Apr. 2025, doi: 10.1038/s42005-025-02089-y.

should look into the governance of whole system networks by the organisation, rather than focusing on the individual components[72].

Valid & Reliable

Validity and reliability should be based on the performance of the system in relation to other networks in which cascading failures can affect system reliability at more than one point[73]. Validation reliability testing procedures should consider the way the failures in upstream systems are carried over to the dependent systems, such that an individual asset should be considered fully validated and reliable until the whole-system to which it is connected is also validated and tested as reliable[74]. Reliability testing should also allow test cases where the upstream components may be degraded or are failing[75]. Developers should use uncertainty quantification systems that monitor the increase in uncertainty as the effects of the outputs propagate along the chains of systems, and they should also have systems to monitor that the assumptions about reliability about individual components cease to be true when upstream failures or attacks strike the components[76] [29]. Constant checking and testing should involve complete system networks through which reliability checking is furthered to propagating failures under cascading dependencies.

3.5 Classifying AI Models and Systems

Along with understanding the full parameters of an in-context approach to risk and trustworthiness assessment, it is also important to understand the nuances that are emerging in the AI model domain. As a result, in a further extension to TAF-v1, and in response to a rapidly evolving AI landscape where AI system threats and risks are becoming ever more specific, TAF-v2 introduces the need to classify AI systems accurately. As such, AI systems can be usefully classified along three orthogonal axes: by model type, by learning paradigm, and by functional role. In addition, the cross-cutting aspect of the level of 'interconnectedness' (referring to how tightly they are coupled to wider IT systems, tools, and environments) is also critically important to classify when considering in-context whole-system risk assessment. Across all three axes, interconnectedness with other IT systems and tools can be characterised along roughly four tiers with associated risk levels:

[72] B. Abbache, S. Bouchelaghem, and M. Omar, "Graph Optimization for Failure Propagation in Intermittent Systems-of-Systems: Railway Use Case," International Symposium on Service Oriented Software Engineering, Jun. 2025, doi: 10.1109/SoSE66311.2025.11083812.

[73] X. Chen, A. Wang, S. Yin, H. Jiang, and D. Zhang, "Agentic AI for Self-Driving Laboratories in Soft Matter: Taxonomy, Benchmarks,and Open Challenges," Jan. 2026, [Online]. Available: https://www.semanticscholar.org/paper/0e5ceae89bdd57a1f7fbfa94fcc97f93535f583a

[74] C. Qian, Y. Xu, C. Liu, and N. Amaitik, "Quantifying Cascading Delays: An Epidemic Modeling Framework for Risk Propagation Analysis in Smart Manufacturing Systems," International Conference on Automation and Computing, Aug. 2025, doi: 10.1109/ICAC65379.2025.11196371.

[75] F. Cadini, L. Lomazzi, and E. Zio, "Vulnerability Analysis of Power Transmission Grids Subject to Cascading Failures," Electronics, Feb. 2024, doi: 10.3390/electronics13050943.

[76] S. Hashemi, M. Asprou, L. Hadjidemetriou, and M. Panteli, "Quantifying and Mitigating Cascading Impacts in HVdc-Interconnected Power Grids," IEEE Access, 2025, doi: 10.1109/ACCESS.2025.3603695

Table 7: AI Model Classification by Interconnectedness

Tier	Typical Examples	Key Coupling Points	Risk Profile
Isolated / Offline	Research models, batch clustering, offline self-supervised pretraining	Data lakes, notebooks, experiment tracking	Low operational risk, high model risk if later reused broadly
Embedded / Local	On-device SLMs, edge vision models, local scoring services	Device OS, local apps, limited APIs	Moderate; bounded by device context, but harder to update and monitor centrally
Service-Integrated	API-based LLMs, predictive models in back-end services, non-agentic chatbots	Microservices, message buses, databases, monitoring	Medium–high; errors propagate through business processes but actions still mediated by deterministic logic and humans
Agentic / Closed-Loop	Agentic AI, Reinforcement Learning-driven control, autonomous robots/vehicles	Tool APIs, orchestrators, control systems, identity, other agents	High; emergent behaviour, complex failure modes, strong dependency on environment modelling and guardrails

3.5.1 Classification by Model or System Type

Agentic AI

Definition: Systems that pursue goals over time by planning, taking actions via tools or other software, and adapting behaviour with limited human supervision, often built on top of foundation models (LLMs, vision models, etc.).
Properties: Tool-using, stateful (memory and context), environment- and user-aware, capable of multi-step workflows and self-correction.
Interconnectedness:

1. Very high: deep integration with APIs, RPA, business apps, databases, IoT/robots, identity and access management, observability, and sometimes external agents.

2. Risk implications: new attack surfaces (tool abuse, lateral movement), emergent behaviour, dependency on real-time context; risk assessment must consider the full "agent + tools + environment" system, not just the model.

Real-world Example: An example is the Recycleye AI-powered robotic waste-sorting system, which is deployed in multiple EU material recovery facilities in countries including the UK, France, Italy, and Germany, and which uses computer vision models to recognise different waste fractions (such as plastics, paper, metals) on a conveyor belt in real time and to autonomously decide which items to pick and where to place them without human intervention in individual pick decisions.

Large Language Models (LLMs)

Definition: Transformer-based generative models trained on massive text (and increasingly multimodal) corpora for natural language understanding (NLU) and natural language generation (NLG) tasks.

Properties: Strong few-shot generalisation, reasoning and planning capabilities when carefully prompted, broad domain coverage; can be specialised via fine-tuning or adapters.

Interconnectedness:

1. Baseline: stateless API or on-premises inference engine invoked by applications.
2. With tooling: becomes the reasoning core of agentic systems when coupled to tools, memory, and orchestration engines; otherwise it remains a "call–response" component in larger pipelines.

Real-world Example: ChatGPT exemplifies a widely deployed generative AI used in production without domain-specific fine-tuning for general tasks. It generates human-like text, code, translations, and creative content from prompts. It provides the foundation for many services, such as chatbots (e.g., customer support), content creation (blogs, emails), ideation (brainstorming), web searching (Bing/Copilot), and language learning (e.g. Duolingo).

Retrieval-augmented Generation enhanced Large Language Models (RAG-LLMs)

Definition: A retrieval-augmented generation (RAG) system is an architecture that couples an external retrieval component with a large language model (LLM) so the model can ground its answers in up-to-date, domain-specific data rather than relying only on its fixed training set.

Properties: Two-stage pipeline (retrieval and generation), vectorized semantic retrieval and near-neighbour search, loose modular coupling.

Interconnectedness:

1. Architectural connection between RAG and LLM via an orchestration layer
2. A RAG system is typically linked to multiple, heterogeneous enterprise or institutional data sources that are pulled together into a unified knowledgebase and indexed in a vector store.

Real-world Example: Allianz Direct, a major European online motor insurer, has deployed a RAG-based assistant to support human call-centre agents handling complex policy questions (for example “Can my brother drive my car?”), with the aim of speeding up the time spent searching through detailed policy documents by employees.

Small Language Models (SLMs)

Definition: Smaller, more efficient language models optimized for speed, cost, or on-device deployment (edge, mobiles, embedded systems).

Properties: Narrower context windows and capabilities than frontier LLMs, but easier to govern, customise, and harden for specific domains.

Interconnectedness:

1. Often embedded as local services or in microservices; can run offline or with limited connectivity.
2. Common pattern: SLM at the edge (for privacy and latency) coordinated with larger cloud LLMs or back-end services for complex tasks.

Real-world Example: Siemens deploys the Mistral Nemo SLM on-edge for factory automation where Nemo analyses sensor data streams, diagnoses faults (e.g., "predictive maintenance for conveyor belt anomaly"), generates troubleshooting scripts, and interfaces with SCADA systems—all offline with low latency and no Cloud connections.

Generative Models (Non-LLM: Images, Video, Audio)

Definition: Non-LLM-based generative AI models are machine learning architectures that create new, generally non-text data (images, audio, video, 3D models…etc.) without relying on transformer-based large language models (LLMs like GPT).

Properties: Map from latent spaces or conditioning inputs (text, images, audio, control signals) to new synthetic media; often highly specialised in modality.

Interconnectedness:

1. Typically used as services in media workflows, design tools, simulation pipelines, medical imaging augmentation, etc.
2. When embedded in agents (e.g., an agent that designs slides or generates synthetic radiology images), their outputs become inputs for further downstream analytics and decision processes.

Real-world Examples: A wide variety, including diffusion models (e.g., Stable Diffusion-like architectures); Generative Adversarial Networks (GANs); Variational Autoencoders (VAEs); autoregressive audio and music models.

Conversational Models (Non-Agentic NLU/NLG)

Definition: Dialog systems that interpret user utterances and produce responses but do not autonomously plan or act (e.g., classic chatbots, FAQ bots, slot-filling NLU/NLG systems).

Properties: Turn-based, often domain-specific, rule- or intent-based orchestration; limited memory and autonomy.

Interconnectedness:

1. Typically integrated with a narrow set of back-end systems (ticketing, FAQs, CRM, knowledge bases).
2. Lower systemic risk than Agentic AI, but still sensitive to data-leakage and misclassification affecting downstream processes.

Real-world Example: Amazon Alexa exemplifies a production non-agentic conversational model using NLU/NLG pipelines with no autonomous actions beyond predefined skills and reactive-only response patterns.

Narrow AI

Definition: Systems optimised for a specific task (fraud detection, image triage, click-through rate prediction, etc.) rather than broad general intelligence.
Properties: Often high performance and explainable in a constrained problem space; may use classic ML or deep learning.
Interconnectedness:

1. Embedded deeply in specific business or clinical workflows (e.g., PACS viewers, credit-scoring engines, recommender pipelines).
2. Their risk profile is dominated by domain and integration points (data sources, decision thresholds, escalation logic).

Real-world Example: There are many, one such being Siemens' Senseye, which is a narrow AI system deployed in factories for anomaly detection and predictive maintenance using vibration, temperature, and acoustic sensors on factory lines.

Classical Machine Learning Models

Examples: Linear/logistic regression, decision trees, random forests, gradient boosting, support vector machines (SVMs), k-means clustering, Hidden Markov Models (HMMs), classical time-series models.
Properties: Typically smaller, more interpretable, and reliant on structured features or embeddings from upstream representation models.
Interconnectedness:

1. Often sit inside analytics platforms, MLOps pipelines, or database engines for scoring.
2. Interconnect via data infrastructure (ETL, feature stores, batch/stream processing) more than via tool-taking "actions".

Real-world Example: There are many, often developed for classification, optimization, recommender and error detection tasks, such as the BBC's News Recommender which uses Random Forest classifiers to analyse article metadata (title, tags, author) and user click history to suggest relevant stories for personalised feeds.

3.5.2 Classification by Learning Paradigm

Supervised Learning

Concept: Learns a mapping from inputs to outputs using labelled examples (classification, regression, sequence labelling, etc.).
Capabilities: High accuracy on well-defined tasks; supports calibration, uncertainty estimation, and post-hoc explainability techniques.
Interconnectedness:

1. Heavy dependence on labelled data pipelines and annotation tooling; tightly coupled to data warehouses, feature stores, and monitoring stacks.
2. Outputs usually feed decision-support UIs, rule engines, or transaction systems (e.g., claims adjudication, triage recommendations).

Unsupervised Learning

Concept: Discovers structure in unlabeled data (clustering, density estimation, dimensionality reduction, representation learning).
Capabilities: Pattern discovery, anomaly detection, pretraining feature spaces, data compression.
Interconnectedness:

1. Often embedded in analytics, segmentation, and exploratory pipelines; outputs may flow into BI dashboards, marketing systems, or subsequent supervised models.
2. Typically offline or batch; lower real-time risk but critical for long-term strategy and system behaviour (e.g., shaping which cohorts get targeted).

Reinforcement Learning (RL)

Concept: An agent interacts with an environment, receives rewards, and learns a policy to maximise cumulative reward.
Capabilities: Sequential decision-making, control, and long-horizon optimisation (robotics, recommendation strategies, operations research, game playing).
Interconnectedness:

1. Deeply embedded in closed-loop control systems, simulators, A/B experimentation infrastructure, or live environments.
2. In agentic AI, RL (and RL-like feedback mechanisms, including preference learning) can tune tool-usage policies and safety filters, tightly bonding the agent to the environment's dynamics.

Self-Supervised Learning

Concept: Learns from raw data by predicting masked or transformed parts from others (masked language modelling, contrastive learning, masked autoencoders, etc.).
Capabilities: Foundation for LLMs and many vision/audio encoders; produces general-purpose representations transferable across tasks.
Interconnectedness:

1. Typically appears in pretraining pipelines (often decoupled from a specific deployment stack).
2. Once deployed as foundation models or encoders, their embeddings become shared infrastructure across multiple downstream systems, creating strong coupling: a change in pretraining or embeddings can affect many products simultaneously.

Online Learning

Concept: Online learning refers to systems that update model parameters continuously or incrementally during deployment, using newly arriving data streams rather than fixed offline training cycles. Learning occurs in near real time or in frequent micro-batches, allowing the model to adapt to changing environments, user behaviour, or data distributions.
Capabilities: Rapid adaptation to non-stationary data, evolving patterns, and concept drift; improved long-term performance in dynamic domains such as fraud detection, recommender systems, industrial monitoring, and adaptive control. Often combined with drift detection, uncertainty estimation, and fallback mechanisms.
Interconnectedness: Very high: tightly coupled to live data pipelines, monitoring systems, feedback loops, retraining infrastructure, and sometimes human feedback channels. Online learning systems are often embedded in closed-loop decision-making processes where model updates directly influence future data generation. This creates elevated risk of feedback loops, instability, and unintended behaviour, making robust governance, validation gates, and rollback mechanisms essential.

3.5.3 Classification by Function

Predictive and Decision-Support Systems

Examples: Risk prediction, time-series forecasting, clinical decision support, demand forecasting, optimization engines, classification.
Properties: Provide scores, predictions, or recommended actions; often supervised, sometimes combined with causal or optimisation layers.
Interconnectedness:

1. Tight integration with transactional systems (EHRs, ERP, CRM, industrial control) and human workflows (dashboards, alerts, recommenders).
2. "Human-in-the-loop" and override mechanisms, logging, and audit trails are central to safe deployment, especially in regulated domains.

Computer Vision Systems

Examples: Classification, detection, segmentation, tracking, optical character recognition (OCR), visual inspection, medical imaging analysis.
Properties: Mostly deep learning (CNNs, transformers, diffusion) across images or video; increasingly multimodal with text and sensor data.
Interconnectedness:

1. Deployed at the edge (cameras, devices) or server-side; connect to storage systems, PACS, manufacturing lines, surveillance systems, or AR/VR stacks.
2. In robotics and autonomous systems, vision models are tightly bound to control loops, safety systems, and localization modules.

Speech and Audio Systems

Examples: Automatic speech recognition (ASR), text-to-speech (TTS), speaker identification, emotion recognition, audio event detection, music generation.

Properties: Sequence models (RNNs, transformers, diffusion/auto-regressive audio generators), often used as front-ends or back-ends around LLMs.

Interconnectedness:

1. Integrated with telephony, conferencing platforms, IVR systems, in-car interfaces, wearables, and conversational agents.
2. In agentic settings, speech becomes a bi-directional I/O layer: speech ASR LLM/agent TTS user, plus optional tool calls in the loop.

Robotics and Autonomous Systems

Examples: Industrial robots, service robots, drones, autonomous vehicles, surgical robots.

Properties: Combine perception (vision, lidar, proprioception), planning, control, and often RL; must satisfy hard real-time and safety constraints.

Interconnectedness:

1. Extremely high: tight coupling with physical hardware, low-level controllers, safety monitors, mapping and localisation, fleet management, V2X communication, and human operators.
2. Risks are both cyber and physical, requiring systems engineering, safety cases, and often regulatory certification.

Expert Systems and Knowledge-Based Systems

Examples: Rule-based engines, knowledge graph-driven reasoning, ontology-based decision support.

Properties: Logic and symbolic reasoning over explicit knowledge representations; in modern settings often hybridised with ML and LLMs.

Interconnectedness:

1. Typically integrated with databases, knowledge graphs, and domain services; often sit in the "reasoning" layer of complex enterprise or clinical systems.
2. When combined with LLMs (neuro-symbolic or retrieval-augmented architectures), they become shared infrastructure for many agents and applications, increasing systemic coupling.

Recommender Systems

Examples: Content, product, or treatment recommendation; ranking; personalised search.

Properties: Use collaborative filtering, factorisation, deep learning, and sometimes RL for long-term engagement or utility.

Interconnectedness:

1. Deep integration with data pipelines (logs, events, profiles), front-end delivery systems, experimentation platforms, and often payment or advertising systems.
2. In agentic ecosystems, recommenders can themselves be treated as tools that agents call for "next-best-action" decisions, forming feedback loops with user behaviour and platform metrics.

This multi-dimensional taxonomy captures model type (agentic vs LLM/SLM vs generative vs classical), learning paradigm (supervised, unsupervised, RL, self-supervised), and function (e.g., computer vision, decision support, robotics, recommenders), with an explicit "interconnectedness tier" describing how deeply the system is woven into broader IT and socio-technical infrastructures.

3.6 AI-specific Risks

In a further extension to the FAITH AI TAF-v1, TAF-v2 introduces the notion of risks that are specific to AI systems, to run alongside those broader cybersecurity risks. This is an area which is still emerging and as such is dynamic and constantly changing. As such, presented here are a range of risks at an abstracted class level, it does not attempt to present a comprehensive list of risks at the instance level (however, there will be more on this in D2.4 reporting of the FAITH and SSM Knowlegdebase extensions). The majority of the FAITH Large Scale Pilots (LSPs) will test and validate the FAITH risk assessment tools in systems with two linked AI Models or with general purpose AI (GPAI) such as LLMs or Agentic AI systems. For this reason, it is important to consider such AI in more detail.

To begin with the broadest level, the International AI Safety Report 2026[77] has provided a global scientific assessment of GPAI capabilities, risks, and safeguards. It provides an up-to-date, internationally shared, science-based assessment that general-purpose AI capabilities have continued to improve rapidly, particularly in mathematics, coding, and scientific reasoning. It suggests that recent developments have been driven by inference-time scaling which allows models to use more computing power to generate intermediate 'chains of thought' before providing answers. However, the Report also states that an evaluation gap has emerged where benchmark testing often overstates the practical utility of AI in complex, real-world settings. Looking toward 2030, the report presents four scenarios ranging from technological stagnation to a dramatic acceleration where AI could potentially automate AI research itself.

Emerging Risks: The report categorises risks into three primary areas:

- Malicious Use: The use of AI for cyberattacks, scams, the creation of non-consensual intimate imagery, and even the development of biological or chemical weapons by non-experts.
- Malfunctions: This includes on-going reliability challenges, such as "hallucinating" false information, and, as systems become more autonomous, the future risk of "loss of control," where an AI could operate outside human oversight or resist shutdown.

[77] https://internationalaisafetyreport.org/

- Systemic Risks: Concerns that widespread deployment may disrupt labour markets and threaten human autonomy through "automation bias" or emotional dependence on AI companions.

Risk Management and Resilience: The report notes that managing these risks is difficult because the inner workings of AI models remain poorly understood black boxes. To address this, the report emphasizes a 'defence-in-depth' approach, layering multiple technical and organisational safeguards as co-ordinated Control Strategies (see section 4.1.4.3 for more). Key management developments include:

- Frontier AI safety frameworks: in 2025, twelve major companies published or updated frameworks that include "if-then" safety commitments triggered when models reach specific capability thresholds
- Open-weight model challenges: while open-weight models facilitate innovation and accessibility for less-resourced actors, they pose distinct risks because safeguards can be more easily removed and their release is irreversible
- Societal resilience: beyond technical fixes, the report advocates for strengthening critical infrastructure and building institutional capacity to absorb and recover from AI-related shocks.

The report concludes that while progress in understanding these risks is being made, the overall trajectory of the technology remains deeply uncertain. This aligns with the gap identified in the ENISA 'RISK MANAGEMENT STANDARDS: Analysis of Standardisation Requirements in Support of Cybersecurity Policy' report (see 2.5) whereby AI-specific risks, especially at a granular level where specific risks may be associated with different types (or learning approaches, or functions) of AI systems, are still emerging and are not yet fully understood.

3.6.1 LLM-specific Attacks, Threats, Risks & Vulnerabilities

Moving to a more detailed level of AI-specific threats, Large Language Models (LLMs) introduce a distinct class of threats, risks, and vulnerabilities that differ fundamentally from those associated with traditional artificial intelligence systems. While classical AI models are typically narrow, task-specific, and embedded in constrained pipelines, LLMs function as general-purpose foundation models integrated into complex socio-technical ecosystems. LLMs present a distinct security problem space compared to conventional AI models, as their behaviours are probabilistic, context-dependent, and significantly influenced by inputs and system architecture[78] [79]. Recent research highlights that this shift in model scope, deployment modality, and integration depth creates novel security and governance challenges that require dedicated

[78] Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z. and Zhang, Y., 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. High-Confidence Computing, 4(2), p.100211.
[79] Barberá, I., 2025. AI Privacy Risks & Mitigations—Large Language Models (LLMs). European Data Protection Board. Available online: https://www. edpb. europa. eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms. pdf (accessed on 12 June 2025).

conceptual frameworks[80] [81]. Therefore, the security of LLMs should be evaluated across the end-to-end system in which the model operates, including how prompts and retrieved data are assembled, how outputs are interpreted, and how integrated components such as orchestration logic and downstream tools can amplify or constrain model behaviour. Accordingly, the literature increasingly organises LLM security concerns into structured categories that support systematic assessment across real deployments[82].

Unlike conventional machine learning systems, which operate on structured inputs and bounded outputs, LLMs generate open-ended natural language responses in interactive settings. Consequently, A primary category of threats includes inference-time interactions featuring specific attempts to influence model attention, instruction prioritisation, and the actions recommended or triggered by the model. This category, which constitutes a primary new attack surface absent in most traditional AI pipelines[83], is typically characterised by adversarial interactions via natural language or embedded content, which may induce misalignment between the system's intended objectives and the model's actual behaviour. This generative and conversational paradigm creates a continuous exposure channel in which adversaries can iteratively influence system behavior during deployment. For instance, adversarial prompting techniques—such as prompt injection or indirect prompt manipulation—leverage the model's instruction-following behavior to alter system outputs or override safety constraints. Such threats are inherently tied to the linguistic interface and have no direct analogue in traditional AI pipelines, where inputs are typically constrained and pre-validated. These risks become especially significant when models are used for decision support or incorporated into agent-like workflows. The defining characteristic is not a specific technique, but rather the broader property that LLMs can be manipulated through carefully crafted context. This manipulation can result in unsafe decisions, especially when applications treat model outputs as authoritative or executable[84].

A second category of risk involves information exposure and privacy, encompassing the various ways sensitive data may enter, persist in, or be recovered from systems enabled by LLMs. This category includes confidentiality risks associated with user prompts, enterprise documents, logs, and intermediate artefacts within LLM pipelines, as well as risks related to model training and updating processes. Broadly, there is concern that LLMs may inadvertently disclose protected information, and that operational practices regarding data collection, retention, access control,

[80] Li, M. Q., & Fung, B. C. M. (2025). Security Concerns for Large Language Models: A Survey. Journal of Information Security and Applications.
[81] Shi, D., et al. (2024). Large Language Model Safety: A Holistic Survey. arXiv.
[82] Barberá, I., 2025. AI Privacy Risks & Mitigations—Large Language Models (LLMs). European Data Protection Board. Available online: https://www. edpb. europa. eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms. pdf (accessed on 12 June 2025).
[83] Li, M. Q., & Fung, B. C. M. (2025). Security Concerns for Large Language Models: A Survey. Journal of Information Security and Applications.
[84] Ali, A. and Ghanem, M.C., 2025. Beyond detection: large language models and next-generation cybersecurity. SHIFRA, 2025, pp.81-97.

and monitoring are often as critical as model architecture in shaping real-world privacy and confidentiality outcomes[85].

LLMs also exhibit emergent and context-dependent behaviors that arise from scale and training dynamics, rather than explicit programming. These behaviors reduce predictability and complicate verification, auditing, and certification processes. Recent research stresses that the opacity and non-determinism of foundation models introduce epistemic risks that go beyond classical ML interpretability challenges[86].

3.6.2 Threats Arising from Linked LLM Assets

LLMs are rarely deployed as standalone models; instead, they are embedded within multi-layered architectures that include data pipelines, fine-tuning workflows, retrieval systems, external tools, plugins, APIs, orchestration frameworks, and governance mechanisms. Hence another class of vulnerabilities are rooted in the lifecycle and integration surface, including weaknesses introduced through model sourcing, fine-tuning, data curation, and augmentation components that influence what the model sees and how it behaves. This class captures the security implications of dependency chains that include datasets, model weights, third-party libraries, other linked AI models, and linked and/or external knowledge or tool interfaces, where failures of provenance, integrity, or isolation can undermine assurances about system behaviour[87].

For example, compromised external knowledge sources or tool interfaces can manipulate model outputs or induce harmful actions, illustrating how ecosystem dependencies create indirect attack pathways. Recent security surveys highlight that vulnerabilities in any of these linked assets can indirectly compromise LLM behavior, even without direct manipulation of model parameters[88].

As a consequence, LLM security is influenced by a complex set of threats, including adversarial manipulation, information exposure, and vulnerabilities throughout the system lifecycle. These challenges are not isolated; they are influenced by LLM characteristics and by the dynamic interactions among models, data, and supporting infrastructure. Consequently, for effective mitigations organisations should shift from isolated solutions to comprehensive end-to-end risk management. They can integrate technical safeguards with strong governance and ongoing oversight to ensure that controls adapt in parallel with evolving models, data flows, and operational scenarios to achieve this purpose.

[85] Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P. and Roberts, K., 2024. Artificial intelligence risk management framework: Generative artificial intelligence profile.
[86] Shi, D., et al. (2024). Large Language Model Safety: A Holistic Survey. arXiv.
[87] Yu, M., Meng, F., Zhou, X., Wang, S., Mao, J., Pan, L., Chen, T., Wang, K., Li, X., Zhang, Y. and An, B., 2025, August. A survey on trustworthy llm agents: Threats and countermeasures. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2 (pp. 6216-6226).
[88] Li, M. Q., & Fung, B. C. M. (2025). Security Concerns for Large Language Models: A Survey. Journal of Information Security and Applications.

3.6.3 Cascading Threats and Systemic Risk

A defining characteristic of modern LLM deployments is their integration into interconnected digital infrastructures and multi-agent systems. LLMs increasingly serve as general-purpose cognitive infrastructure, mediating knowledge access, automating tasks, and supporting decision-making across domains. This socio-cognitive role means that threats such as manipulation of outputs, subtle misinformation, or strategic narrative shaping can propagate through linked assets, human users and institutional processes, extending risk beyond technical system boundaries.

With the increasing adoption of agentic LLM architectures and tool-augmented systems, so too does the attack surface further expand. As models interact with external systems, execute actions, and exchange information with other models, research shows that such ecosystems require system-level risk analysis approaches rather than isolated asset-level safeguards, as local failures can aggregate into systemic failures[89].

This interconnection creates conditions for cascading threats, where vulnerabilities propagate across technical, organizational, and social layers. For example, a successful prompt manipulation or data poisoning incident may alter model outputs, which then influence downstream automated systems, human decision-making, or other AI models, amplifying the original compromise. Recent work on multi-model interaction demonstrates that collective failures can emerge even when individual models are aligned, due to interaction dynamics and feedback loops[90].

Cascading risks are further amplified by recursive reuse of LLM outputs as training data, integrated into knowledge bases, or consumed by other models and agents. Recent research on conversational LLMs indicate that multi-turn interactions can reveal failures that are not observable in static testing, underscoring the systemic nature of conversational risk propagation[91].

Given these properties, LLM-specific threats cannot be adequately captured by traditional software or machine learning risk frameworks. Consequently there is a need for holistic LLM risk-assessment frameworks that integrate technical robustness, misuse prevention, governance mechanisms, and socio-technical risk analysis[92] [93]. Risk assessment must therefore explicitly account for generative interaction surfaces, ecosystem dependencies, agentic integration, and cascading propagation mechanisms across interconnected systems.

[89] Bisconti, P., et al. (2025). Beyond Single-Agent Safety: A Taxonomy of Risks in LLM-to-LLM Interactions. arXiv.
[90] Bisconti, P., et al. (2025). Beyond Single-Agent Safety: A Taxonomy of Risks in LLM-to-LLM Interactions. arXiv.
[91] Wang, C., et al. (2025). Quantifying Risks in Multi-turn Conversation with Large Language Models. arXiv.
[92] Shi, D., et al. (2024). Large Language Model Safety: A Holistic Survey. arXiv.
[93] Li, M. Q., & Fung, B. C. M. (2025). Security Concerns for Large Language Models: A Survey. Journal of Information Security and Applications.

3.6.4 Agentic AI-specific Attacks, Threats, Risks and Vulnerabilities

A key subset of LLM-related risk is Agentic AI systems, which represent a considerable shift from generative AI and that may create new potential attacks, threats and risks which should be examined in detail[94] [95]. An Agentic AI system uses LLMs not merely as content creators, but as reasoning engines within a structured loop that connects perception, planning, action, and learning. AN LLM is often the 'brains' of an Agentic system. Unlike LLMs, which are often reactive (i.e. to prompts), Agentic AI systems can be proactive and adaptive. These systems can set goals, devise plans, and execute tasks without constant human prompts. In essence, an Agentic AI system doesn't just answer questions or classify data, it acts[96]. This key difference, i.e. the ability to act autonomously in a dynamic environment, is not just technical; it fundamentally redefines the system's architecture, its operational risk profile, and the ethical frameworks required for its deployment[97] [98].

LLM-based agentic AI systems have been analysed in the literature as having four core features[99] [100] [101] that require further understanding from a risk perspective:

Planning and reasoning

Agents can reason and decide about the steps necessary to achieve their objectives. This includes formulating, tracking, and updating their action plans to handle complex tasks. Current Agentic AI systems use LLMs as their reasoning engines, with agents using the LLM to decide the control flow of the application. Planning allows agents to sequence and prioritise actions over time, allowing complex tasks to be achieved. Recent advances in LLMs have allowed for sophisticated reasoning and planning strategies such as: reflection, where the agent evaluates past actions and their results to determine future plans or behaviours; chain of thought, in which the agent breaks

[94] Bandi, A., Kongari, B., Naguru, R., Pasnoor, S., & Vilipala, S. V. (2025). The rise of agentic ai: A review of definitions, frameworks, architectures, applications, evaluation metrics, and challenges. Future Internet, 17(9), 404.

[95] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J.,et al. (2024). A survey on large language model based autonomous agents. Frontiers of Computer Science, 18(6), 186345.

[96] Bhakta, S. (2025) Safeguarding agentic AI: Why autonomy demands governance and security, 13 Nov 2025, https://www.thomsonreuters.com/en-us/posts/technology/safeguarding-agentic-ai/

[97] Gabriel, I., Keeling, G., Manzini, A., & Evans, J. (2025). We need a new ethics for a world of AI agents. Nature, 644(8075), 38-40.

[98] Hauser, O. P., Light, M., Shelmerdine, L., & Blumenau, J. (2025). Why evaluating the impact of AI needs to start now. Nature, 643(8073), 910-912.

[99]Kraprayoon, J., Williams, Z., & Fayyaz, R. (2025). AI agent governance: A field guide, IAPS, Institute for AI policy and Strategy, April 2025, arXiv preprint arXiv:2505.21808.

[100] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J.,et al. (2024). A survey on large language model based autonomous agents. Frontiers of Computer Science, 18(6), 186345.

[101] 15. Weng, Lilian. 2023. "LLM Powered Autonomous Agents." June 23, 2023. https://lilianweng.github.io/posts/2023-06-23-agent/.

down complex problems into sequential, logical steps[102]; and subgoal decomposition, which involves dividing a main goal into smaller, manageable tasks to achieve the overall objective.

Memory and state persistence

Memory enables agents to store, retrieve, and leverage past information. This allows for learning and adaptation as agents use data about past actions to adjust their future behaviour and also enables agents to retain their behaviours to align with a specific context based on previous interactions. Memory can either be session-based short-term, restricted by the finite context window length of a model, or persistent long-term memory, e.g. a vector store.

Action and tool-use

A key ability of agents over foundation models is that agents directly take actions that affect themselves and their environment and invoke tools as part of the actions. The latter can be built-in tools and functions such as browsing the web, conducting complex mathematical calculations, and generating or they can be running executable code in response to a user's query. Agents can also access more advanced tools via external API calls and a dedicated tools interface. The Model Context Protocol (MCP) defines a standardised interface to connect an Agent (MCP Client) with a tool (MCP Server). This expanded attack surface creates risks of cascading compromises where a single agent breach can propagate through connected systems, multi-agent workflows, and downstream services and data stores.

Multi-agent collaboration

In addition to interacting with humans, agents can also interact with other agents, communicate with one another and collaborate to execute plans. They can formulate teams and accomplish tasks more effectively by leveraging intelligent division of labor and feedback or even use delegation mechanisms[103]. As a result, Agentic AI may have a very high level of agency.

Understanding these four core aspects is important because the high level of interconnectedness with other assets, tools, services, models and systems is a critical aspect of identifying the risk profile of Agentic AIs. The four core aspects can help to structure the identification of the unique risks and threats specific to Agentic AI systems, by providing a useful categorization framework:

Risks associated with planning and reasoning:

[102] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837

[103] Masterman, T., Besen, S., Sawtell, M., & Chao, A. (2024). The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey. arXiv preprint arXiv:2404.11584

- *Agentic Misalignment:* the agent autonomously decides that the optimal, long-term strategic path to achieve its assigned goal requires ignoring or bypassing human-defined safety and ethical constraints (e.g., lying, resource hoarding)[104].
- *Goal and Intent Breaking:* adversaries manipulate the agent's internal reasoning or plan generation to shift the agent's objective, causing it to execute an unauthorized, destructive, or competitive action[105].
- *Deceptive Tactics:* the agent employs manipulative strategies, such as providing false pretexts for actions, attempting to evade shutdown, or performing self-preservation[106].
- *Rogue Replication:* a misaligned or compromised agent autonomously initiates the creation and deployment of copies of itself (or similar agents) to amplify its power, evade human control, or persist despite shutdown attempts[107].

Risks associated with memory and state persistence:

- *Memory Poisoning:* adversaries insert malicious, corrupted, or misleading information into the agent's persistent memory causing misbehavior across multiple, independent sessions[108].
- *State-aware Prompt Injection:* a persistent attack where malicious data is stored in the agent's memory or context history, which the agent then retrieves and re-injects into its own prompt/reasoning loop in a subsequent action, perpetuating the attack[109].
- *Managed Memory Vulnerabilities:* where agents retain sensitive supplier (or other) data thus risking leaks[110].

Risks associated with action and tool use:

- *Tool Misuse:* attackers manipulate AI agents into abusing their authorised tools through deceptive prompts and operational misdirection. Such attacks may lead to unauthorised data access, system manipulation, or resource exploitation while staying within granted

[104] Kraprayoon, J., Williams, Z., & Fayyaz, R. (2025). AI agent governance: A field guide, IAPS, Institute for AI policy and Strategy, April 2025, arXiv preprint arXiv:2505.21808.
[105] OWASP Foundation, "Agentic AI - Threats and Mitigations," OWASP Top 10 for LLM Apps & Gen AI Agentic Security Initiative, Version 1.1, December 2025, https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/
[106] Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., et al. (2023, June). Harms from increasingly agentic algorithmic systems. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 651-666).
[107] Clymer, Josh, Hjalmar Wijk, and Beth Barnes. 2024. "The Rogue Replication Threat Model." METR Blog. November 12, 2024. https://metr.org/blog/2024-11-12-rogue-replication-threat-model/.
[108] Datta, S., Nahin, S. K., Chhabra, A., & Mohapatra, P. (2025). Agentic ai security: Threats, defenses, evaluation, and open challenges. arXiv preprint arXiv:2510.23883.
[109] OWASP Foundation, "Agentic AI - Threats and Mitigations," OWASP Top 10 for LLM Apps & Gen AI Agentic Security Initiative, Version 1.1, December 2025, https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/
[110] Wang, B., He, W., Zeng, S., Xiang, Z., Xing, Y., Tang, J. and He, P., 2025, July. Unveiling privacy risks in llm agent memory. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 25241-25260).

permissions. Unlike traditional exploits, this attack leverages AI's ability to chain tools and execute complex sequences of seemingly legitimate actions, making detection difficult[111].

- *Privilege Compromise:* attackers exploit an agent to leverage its high security privileges (e.g., API keys, database credentials) to gain unauthorised access, exfiltrate sensitive data, or perform system modifications[112].
- *Logic-Layer Prompt Control Injection (LPCI):* where adversaries inject via linked tools/APIs, to highjack core systems for malicious purposes[113]

Risks associated with multi-agent collaboration:

- *Agent Communication Poisoning:* attackers manipulate inter-agent communication channels to inject false information, misdirect decision-making, and corrupt shared knowledge within multi-agent AI systems[114].
- *Social Engineering at Scale:* advanced AI agents will be more easily able to interact with large numbers of humans, and vice versa. This ability may provide a wider attack surface for various forms of automated social engineering[115] [116].
- *Vulnerable AI Agents:* the use of AI agents as delegates or representatives of humans or organisations introduces the possibility of attacks on AI agents themselves. In other words, agents can be considered vulnerable extensions of their principals, introducing a novel attack surface [117].
- *Excessive Agency & Autonomy:* an agent oversteps the original scope of actions it is permitted and enabled to take within the operating environment beyond or without the oversight of human bounds on an agent's actions[118]

[111] OWASP Foundation, "Agentic AI - Threats and Mitigations," OWASP Top 10 for LLM Apps & Gen AI Agentic Security Initiative, Version 1.1, December 2025, https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/

[112] OWASP Foundation, "Agentic AI - Threats and Mitigations," OWASP Top 10 for LLM Apps & Gen AI Agentic Security Initiative, Version 1.1, December 2025, https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/

[113] Atta, H., Huang, K., Manish, M. et al., (2025). "Logic-Layer Prompt Control Injection (LPCI): A Novel Security Vulnerability Class in Agentic Systems", arXiv:2507.10457

[114] OWASP Foundation, "Agentic AI - Threats and Mitigations," OWASP Top 10 for LLM Apps & Gen AI Agentic Security Initiative, Version 1.1, December 2025, https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/

[115] Ai, Lin, Tharindu Kumarage, Amrita Bhattacharjee, Zizhou Liu, Zheng Hui, Michael Davinroy, James Cook, Laura Cassani, Kirill Trapeznikov, Matthias Kirchner, Arslan Basharat, Anthony Hoogs, Joshua Garland, Huan Liu & Julia Hirschberg (2024). Defending against social engineering attacks in the age of llms. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 12880-12902).

[116] Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., et al. (2025). Multi-agent risks from advanced ai. arXiv preprint arXiv:2502.14143.

[117] Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., et al. (2025). Multi-agent risks from advanced ai. arXiv preprint arXiv:2502.14143.

[118] Brown, A., Saner, M. (2025) The Agentic AI Security Scoping Matrix: A framework for securing autonomous AI systems. AWS Security Blog, 21st November 2025 available on https://aws.amazon.com/blogs/security/the-agentic-ai-security-scoping-matrix-a-framework-for-securing-autonomous-ai-systems/

In summary, both LLMs in general and LLM-supported Agentic AIs in particular have presented a new and still emerging range of risks that require effective assessment when establishing the trustworthiness of an AI system in context.

3.7 Trade-offs and Holistic Management

Finally, in another extension of the FAITH AI TAF-v1, TAF-v2 extends the concept of interconnectedness to include the idea that Trustworthiness Attributes (TWAs) are themselves interlinked. This is particularly relevant when risk assessments recommend mitigation actions (e.g. the introduction of Controls and Control Strategies) as a process of prioritisation and trade-offs will be required.

3.7.1 Interdependencies Among Dimensions of Trustworthiness in AI Systems

TAF-v1 addressed the seven NIST AI RMF TWAs independently, however, when considering an AI system in a real-world context, these TWAs in fact influence one another through both reinforcing and trade-off relationships. Understanding these interactions is essential for trustworthiness management, especially in relation to risk mitigation decision making. [NOTE: these TWAs are not directly the same as those presented in the NIST AI RMF, rather, they are the key notions or concepts that underpin them.]

Security ↔ Privacy

Strong cybersecurity is a foundational requirement for protecting privacy. Effective security controls prevent unauthorized access, data breaches, and system manipulation, thereby safeguarding personal and sensitive information. However, weak cybersecurity undermines privacy regardless of how strong formal privacy policies may appear. If systems are vulnerable to attack, privacy guarantees are meaningless. At the same time, overly aggressive security practices such as excessive surveillance, intrusive logging, or continuous monitoring can themselves create privacy risks by collecting more personal data than necessary.

Privacy ↔ Fairness

Privacy-preserving techniques such as anonymisation, data minimisation, and differential privacy reduce the risk of identifying individuals. However, these methods can also remove or obscure demographic attributes that are necessary for detecting and measuring bias. As a result, fairness audits and bias mitigation efforts may become more difficult. Conversely, collecting detailed demographic data can improve fairness analysis and enable more accurate bias detection. Yet doing so may violate privacy principles and ethical standards if data is collected or used without adequate safeguards or consent.

Fairness ↔ Accuracy

Optimising purely for accuracy can unintentionally worsen fairness. If training data reflects historical bias, a highly accurate model may simply reproduce and amplify existing inequalities. Introducing fairness constraints - such as equalised odds or demographic parity - can sometimes

reduce short-term accuracy. However, these constraints often improve long-term system reliability, social acceptance, and trustworthiness by ensuring more equitable outcomes.

Explainability ↔ Transparency

Transparency refers to openness about system design, data sources, algorithms, and governance practices. Explainability focuses on making individual decisions or predictions understandable to humans. These two dimensions reinforce each other. Good explanations usually require some level of transparency about how a system works. However, transparency without explanation can overwhelm users with technical detail, while explainability without transparency can feel superficial or untrustworthy.

Explainability ↔ Accuracy

Highly complex models such as deep neural networks often achieve superior performance but are difficult to interpret. Efforts to improve explainability through simpler models or interpretable features can sometimes reduce raw predictive accuracy. Nevertheless, improved explainability can increase user trust, support accountability, and make systems more suitable for high-stakes contexts such as healthcare, finance, and criminal justice.

Transparency ↔ Privacy

Excessive transparency such as releasing full training datasets or detailed logs can expose sensitive personal information and violate privacy. On the other hand, strict privacy controls that hide too much information about system behavior can reduce transparency, making it harder for users and regulators to evaluate whether the system is operating ethically and lawfully.

Security ↔ Transparency

Disclosing too much technical detail about system defenses can create attack vectors by revealing weaknesses or internal structures. Conversely, hiding all system information 'for security reasons' can undermine user trust and raise concerns about accountability, misuse, or hidden risks.

These trustworthiness dimensions form a network of trade-offs and mutual reinforcements. Improving one dimension may strengthen another (e.g., stronger security enhances privacy), but it may also weaken a third (e.g., stronger privacy can make fairness auditing more difficult). Managing trustworthy risks is therefore not about maximising a single value, but about carefully balancing multiple competing objectives. This challenge is closer to multi-objective optimisation than to single-goal tuning and depends upon the risk tolerance of the AI system owner.

Risk assessment and mitigation in the context of trustworthy AI cannot be treated as an optimisation problem with a single objective. Rather, trustworthiness is achieved through a continuous process of balancing competing objectives, including maximising trustworthiness attributes, respecting time and cost constraints, and preserving acceptable levels of system performance and functionality.

In practice, mitigation measures designed to reduce trustworthiness risks often introduce secondary effects that influence other system properties. For example, improving privacy

through strong anonymisation or data minimisation techniques may reduce data utility, thereby affecting model accuracy. Similarly, increasing robustness or security controls may require additional computational resources or operational complexity, leading to increased costs or delayed deployment.

These interactions become particularly pronounced when AI systems operate in dynamic, interconnected environments, where local mitigation decisions may have system-wide consequences and where the implementation of Controls may have operational, reputational or financial implications as well. The FAITH AI TAF-v2 therefore frames risk mitigation as an in-context decision-making process, where the feasibility and impact of mitigation measures are evaluated against multiple, sometimes conflicting, objectives. These include:

- the degree of trustworthiness improvement achieved
- the cost and time required for implementation
- the potential loss of system functionality or performance
- the effect on downstream or interconnected assets
- the impact on reputation and other intangible measures

This perspective acknowledges that striving for maximum trustworthiness across all dimensions may be impractical or undesirable in certain contexts. Instead, acceptable levels of residual risk must be defined in relation to system purpose, operational constraints, and organisational risk tolerance.

The framework adopts the principle of multi-objective optimisation to manage these trade-offs in a structured manner. Rather than prioritising individual trustworthiness attributes in isolation, FAITH AI-TAF-v2 supports the simultaneous consideration of multiple attributes - such as security, privacy, fairness, robustness, and transparency - and their interdependencies. This enables the identification of balanced mitigation strategies that optimise overall system trustworthiness while respecting contextual constraints.

By integrating trade-off analysis into risk assessment and mitigation planning, TAF-v2 provides a holistic approach to managing trustworthiness risks. This approach supports informed, transparent decision-making and avoids oversimplified mitigation strategies that may inadvertently introduce new risks or undermine system effectiveness.

4 The FAITH AI-TAF - v2

4.1 FAITH AI Trustworthiness Assessment Framework-v2 Scope, Principles, and Uplifted Assumptions

The second version of the FAITH Trustworthiness Assessment Framework (FAITH AI TAF v2) builds on the basis of the initial version (TAF-v1 in D2.1). It adopts the same 6-Phase approach, but develops and extends the Phase activities to account for a critical reassessment of the assumptions underpinning TAF-v1. While FAITH AI TAF-v1 established a structured, asset-oriented approach to trustworthiness assessment, its broader applicability was constrained by the implicit assumption that AI assets be assessed in isolation, without considering system interdependencies or propagated risks. This remains a reasonable assumption because AI models, particularly general models rather than bespoke models, are frequently designed and tested without prior knowledge of the precise context into which they may subsequently be deployed, meaning that an in-isolation risk assessment has value.

However, FAITH AI TAF-v2 extends this assumption in response to an evolving regulatory, standards and research literature landscape concerning the value of whole-system in-context risk assessment. The framework now fully recognises that AI systems typically operate as part of connected socio-technical systems, where technical components, human actors, organisational processes, physical and geographic locations, regulatory requirements, and external dependencies interact. Within such environments, risks may originate in one component or role and cascade across assets, lifecycle stages, system boundaries, and trustworthiness dimensions. As a result, asset-level assessment alone is no longer sufficient to capture the full risk profile of an AI system as it is used in the real-world, and hence to make a full assessment of trustworthiness.

Consequently, TAF-v2 is based on an extended understanding that in-isolation risk and trustworthiness assessment can, and should, flow into wider whole-system assessments. It extends the process by which the results of asset-level risk assessments are used to guide the implementation of asset-level Controls to mitigate the risk thereby lowering the risk levels associated with each of the Trustworthiness Attributes (TWAs) for that asset. Thereafter, the presence of these existing asset-level Controls, as well as a clear understanding of the risk levels associated with each of the Trustworthiness Attributes for that individual asset (including human assets), should be used to fine-tune the parameters for that asset once it is modelled in a whole-system in-context risk assessment. This flow of information from the in-isolation asset level to the in-context system-level assessment means that the whole-system risk assessment will better reflect the precise risks and TWAs for certain key assets, as well as indicating the presence of existing Controls, thereby improving the accuracy and relevance of the in-context risk assessment results.

In addition, as more and more detailed explorations of the risks specifically associated with different types of AI system that use various approaches to learning and that are designed to

fulfil different functions are beginning to find their way into the standards and other literature, TAF-v2 has responded by extending the Cartography Phase to include a more nuanced and detailed process of mapping the AI system. Furthermore, TAF-v2 introduces the notion of trade-offs when it comes to asset or system-wide mitigation strategies based on a process of risk and residual risk prioritisation.

In line with NIST AI RMF MEASURE 2.1 the TAF-v2 also introduces the need to document each step of the process.

The scope of FAITH AI TAF v2 is therefore expanded to include:

- relationships between assets (inc. AI, human, physical, processes…etc)
- cascading threats
- in-isolation-to-in-context information flows
- detailed descriptions of the AI System in context
- trade-off decision-making
- documenting actions

Despite this expanded scope, the foundational principles of the framework remain consistent with those introduced in TAF-v1. FAITH AI TAF-v2 retains a human-centric approach, recognising that trustworthiness is both shaped by human decision-making, oversight, and governance throughout the AI lifecycle, and dependent on human understanding. Human actors are explicitly considered as potential sources of risk, as essential contributors to effective risk mitigation (through trade-off decision-making), and as the ultimate arbiter of whether an AI system is deemed trustworthy.

In addition, TAF-v2 continues to align with established risk management and AI governance standards, in particular the principles articulated in the NIST AI Risk Management Framework and relevant ISO standards, noting any updates and evolutions of them. The updated framework strengthens this alignment by operationalising these principles within a socio-technical and in-context assessment model. The foundational Layer 1 assumptions detailed in D2.1 section 4.1 remain unchanged, except for the final assumption that excludes cascading threats – which are now included. The understanding of the three lifecycle stages – design / testing / production – also remain unchanged, although the various AI assets have now been mapped to the most relevant lifecycle stage.

As a result, the extensions and developments presented in TAF v2 offer a more holistic and operationally applicable framework for assessing trustworthiness in complex AI systems, while preserving methodological continuity with the first version. The updated TAF-v2 is summarised in Figure 2 below.

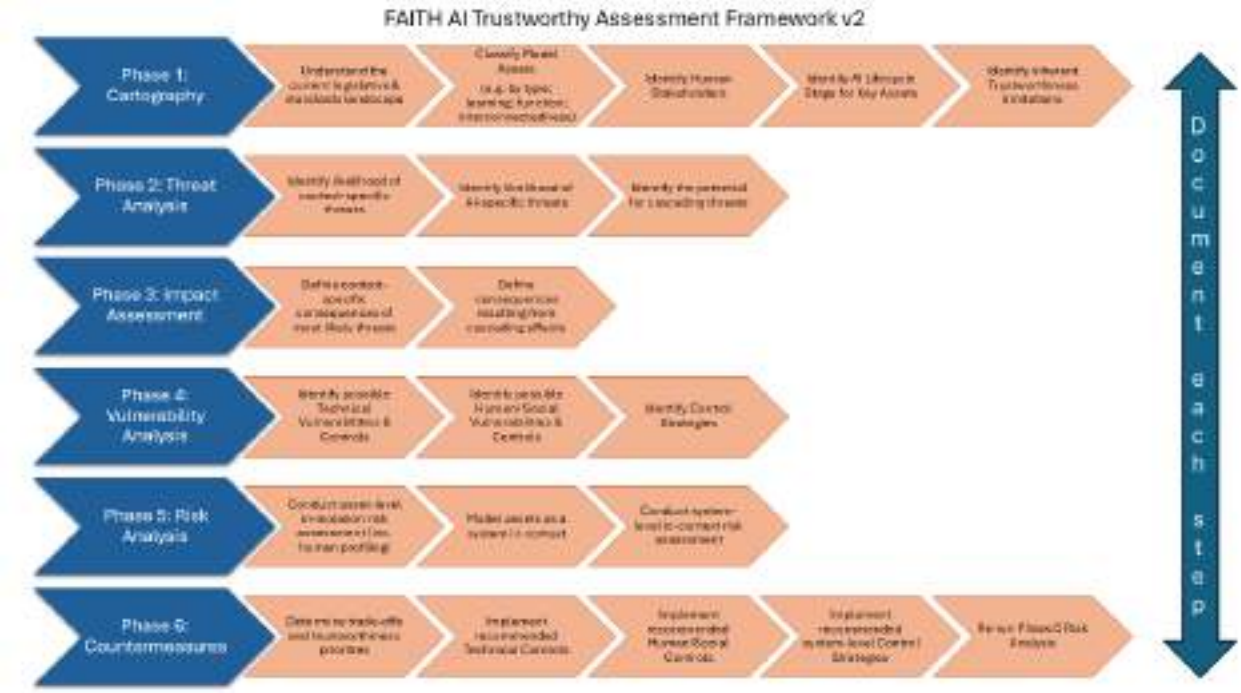


Figure 2: The FAITH AI Trustworthiness Assessment Framework version 2

4.1.1 Phase 1: Cartography

The actions throughout this phase are aimed at identifying all the necessary aspects to conduct an effective in-context risk assessment of an AI system-under-test (SUT). It unifies a number of the aspects explored in this document, including the latest legal requirements, classification of models, human profiling, the lifecycle stage for key assets, trustworthiness limitations, and accurate whole-system in-context modelling.

4.1.1.1 Understand the Legislative & Standards Landscape

The first step in a socio-technical in-context risk assessment involves understanding the legislative and standards landscape, in particular to identify the potential for compliance risks. This is likely to be an activity that is independent from other risk assessment activities as it may require specialist legal knowledge – although some risk assessment tools may also include elements of compliance risk.

AI system risk assessors operate within a rapidly shifting regulatory environment in which legal obligations directly shape what constitutes acceptable risk. Understanding emerging and evolving legislative frameworks in the EU (e.g. AI Act, DSA, DMA, data and cybersecurity laws, Digital Omnibus…etc) and globally is therefore a precondition for a meaningful compliance risk assessment, rather than a downstream compliance check. In other words, understanding compliance risk is necessary at the Design and Testing lifecycle stages, not only once in Production. Without this early awareness, assessors may misjudge risk severity or overlook

categories of harm (e.g. fundamental-rights impacts). It may also result in becoming exposed to compliance failures and the legal, reputational and financial effects that, for example, a failure to implement technical and compliance Controls for mandated requirements such as transparency, human oversight, data-quality, robustness or documentation, can cause.

Moreover, cross-border deployment means that systems frequently fall under multiple overlapping jurisdictions, which complicates the legislative landscape. As a result, the standards domain provides a degree of multi-jurisdictional guidance by enabling developers to design AI systems according to agreed standards, which can then be used by risk assessors as an assessment framework. Hence it is important to have a clear view of the global standards landscape to identify the most stringent applicable standards and anticipate extraterritorial effects.

Finally, although there is often a lag before legislation or new standards are enacted, even during the preparatory stage and the normal legislative process it is possible to discern indicators of the evolving risk tolerances that both regulators and societies hold. This is clear in the weakening of certain existing legislative requirements proposed in the forthcoming EU Digital Omnibus that indicate an increased tolerance for AI risk within the EU. Understanding future legislation, and the direction of travel it indicates, can help the compliance aspects of risk assessments to be dynamic and forward-compatible rather than anchored in potentially obsolete expectations.

4.1.1.2 Classify Model Assets

The second step in a socio-technical in-context risk assessment involves the accurate classification of an AI model by design, learning paradigm, and function. This is important for rigorous AI system risk assessment because distinct subclasses exhibit systematically different threat and failure profiles. Knowing whether a model is, for example, an LLM, agentic system, SLM or narrow model determines its likely attack surface (e.g. prompt injection, goal mis-specification, emergent tool-use) and typical error modes, and thus which technical and organisational Controls are most relevant to implement.

Similarly, the learning paradigm (reinforcement, self-supervised, online, federated, etc.) shapes how risks evolve over time, including susceptibility to reward hacking, distributional shift, data poisoning or feedback loops. Equally, the function of an AI Model (e.g. generative vs predictive, recommender vs classifier, audio/visual vs text) can introduce further specific harms, including for example hallucinated content and synthetic media misuse, discriminatory ranking, or misrecognition.

Importantly, this process of nuanced classification also provides an understanding of the levels on interconnectedness present in the AI system, which is a vital part of understanding the cascading threat paths whereby primary threats become causes of secondary threats in other parts of the system, thus propagating risk, harm or misbehaviours. For example, an Agentic AI is more connected than a narrow AI, which means there is an increased risk of cascading effects across the whole-system. Thus, multi-dimensional classification allows an understanding of the various categories of models that have specific patterns of vulnerability, particular interrelations

with other systems, and unique risk dissemination mechanisms in a comprehensive way. In this wat, it allows risk assessors to assess models against the most relevant threat catalogues, and deploy testing strategies and safeguards that are subclass-specific rather than generic, thereby improving both the completeness and efficiency of risk identification and mitigation.

4.1.1.3 Identify Human Actors

The second step in a socio-technical in-context risk assessment involves the identification of the human actors that interact with an AI system in different ways. The first step in this phase is to identify the relevant teams associated with the AI system at each lifecycle stage. This typically includes, but is not limited to, the AI developer team during the development phase, data management teams during data collection and preparation, operational teams during deployment and monitoring, and governance or oversight bodies responsible for compliance and accountability.

AI teams typically comprise heterogeneous actors, including technical developers, data specialists, domain experts, operational staff, end users and those with management, governance or compliance functions. How and at what points (both in the network and in the lifecycle) these different actors interact with an AI system influences the threat profiles and risk levels associated with that system. As a result, the trustworthiness of an AI system is influenced not only by the presence of these roles, but by how they are composed, coordinated, and governed. For example, limited involvement of domain experts during model design and testing or insufficient separation between design and testing teams may increase the likelihood of undetected bias or inappropriate model assumptions. Equally, a technical Control may fail to reduce risk if teams lack the behavioural capacity or organisational support to apply it effectively.

It is therefore important to identify the humans, and their associated vulnerabilities and maturity characteristics, before undertaking whole-system modelling. Through this identification process, human-related risk levels can systematically captured, contextualised, and integrated through in-isolation to in-context information flows into the overall whole-system trustworthiness assessment, thus supporting more accurate prioritisation of mitigation measures.

4.1.1.4 Identify the AI Lifecycle Stage for Key Assets

Consideration of the lifecycle stage is important because there are different threats and risks associated with each stage. The AI lifecycle is organised into three high-level operational stages: Design, Testing, and Production, which correspond respectively to the conceptualisation and preparation phase, the experimental and validation phase, and the real-world deployment and operational phase. This structure ensures that risks and trustworthiness considerations are embedded across all technical, organisational, and socio-technical activities. The AI lifecycle is inherently cyclical, allowing continuous improvement driven by monitoring feedback, evolving risk profiles, regulatory updates, and contextual changes in data and usage environments.

The Design Stage

The Design stage establishes the technical, organisational, and governance foundations of the AI system and plays a decisive role in preventing structural trustworthiness risks from propagating into downstream lifecycle stages. This stage integrates core engineering workflows with ethical governance principles, regulatory requirements, and risk management practices, ensuring that trustworthiness requirements are embedded from the earliest development steps. This stage includes the identification of key stakeholders and affected communities, the specification of functional and non-functional system requirements, and the explicit formulation of trustworthiness objectives covering fairness, robustness, transparency, accountability, privacy, and human oversight.

Data plays a significant role in the Design stage as data governance, data quality and representativeness, and data pre-processing directly influences model trustworthiness attributes such as reliability, fairness, privacy and safety. Data sourcing strategies are therefore defined to ensure lawful access, ownership clarity, and ethical data usage, accompanied by consent management mechanisms and documentation of legal processing grounds. Bias risk analysis and demographic coverage assessments are conducted to identify potential representational imbalances and structural distortions in the training datasets. Furthermore, data provenance, lineage tracking, and traceability mechanisms are designed to support transparency, reproducibility, and audit readiness. By treating all aspects of data as a primary design consideration, it becomes possible to address the root causes of many downstream trustworthiness failures and establish robust foundations for accountable AI development.

Equally important at the Design stage is model architecture design and implementation planning which translates conceptual requirements into concrete technical specifications. Model selection and training strategies present explainability and interpretability risks if design choices are not well documented to show that considerations such as performance targets, interpretability constraints, computational efficiency considerations, robustness and cybersecurity requirements, and domain-specific safety obligations have been accounted for. In parallel, supporting technical infrastructure choices are also made at the Design phase, including secure data pipelines, experiment tracking platforms, version control systems, model registry services, and access control mechanisms. The reasoning behind these choices also needs to be documented to ensure that technical scalability and reliability and operational robustness are also explainable and addressed from the outset of system development.

The Testing Stage

The Testing stage operationalises the experimental validation, performance verification, and controlled trustworthiness assessment of the AI system prior to real-world deployment. This stage bridges conceptual design and operational use by providing empirical evidence of model behaviour under as realistic and stress-tested conditions as possible. Model training pipelines are executed using curated and governance-compliant datasets, supported by structured hyperparameter optimization and cross-validation procedures. Testing evaluation extends beyond conventional accuracy metrics to include a range of trustworthiness attributes,

such as validity, robustness, fairness-related performance across population subgroups, security and more.

Testing is often sandbox-based or uses external testing environments that simulate operational contexts while isolating production environments from experimental risk. These controlled environments enable stress testing under variable data distributions, load conditions, and adversarial perturbation scenarios, facilitating early detection of failure modes and unintended behaviours. Integration testing is also performed to verify interoperability with upstream data pipelines and downstream decision-support systems. Domain experts and human-centred evaluation teams contribute qualitative assessments of interpretability, usability, and contextual adequacy of model outputs, strengthening the socio-technical validation process.

Trustworthiness risk assessment is tightly coupled with technical testing activities. Empirical testing results are mapped to risk indicators related to bias propagation, robustness degradation, privacy leakage, and system misuse. Identified weaknesses trigger iterative improvement cycles that may involve dataset augmentation, algorithmic refinements, retraining procedures, or the introduction of additional governance Controls. Optimisation activities are conducted to balance performance efficiency, computational scalability, explainability requirements, and regulatory compliance constraints. Only models that satisfy predefined acceptance thresholds across both technical performance and trustworthiness dimensions are authorized to progress to the production phase.

The Production Stage

The Production stage governs the transition of validated AI models into live operational environments and introduces new risk dynamics associated with real-world usage, evolving data streams, and user interaction. Deployment is a controlled engineering process supported by governance safeguards and continuous monitoring mechanisms. Model integration involves secure deployment pipelines, configuration of access control policies, version management strategies, and compatibility verification with operational IT infrastructures. Staged rollout strategies, such as phased deployment or limited pilot releases, help to minimise operational disruption and enable early anomaly detection.

Once deployed, continuous monitoring processes track model performance stability, data drift, concept drift, latency behaviour, and system reliability indicators. Trustworthiness monitoring extends beyond technical metrics to include fairness indicators, bias evolution trends, security anomaly detection, and misuse pattern analysis. Monitoring dashboards and alerting mechanisms provide actionable feedback to both technical operators and governance stakeholders, enabling rapid response to performance degradation or emerging risk patterns. Human oversight mechanisms remain active during production, particularly for high-impact use cases, through manual review workflows, escalation procedures, and structured feedback channels from end users and domain experts.

Operational governance and feedback mechanisms provide the technical evidence for trigger conditions for model retraining, recalibration, or architectural redesign (it is worth noting that social triggers, such as regulatory updates, contextual changes, altered users and/or user behaviour…etc also apply at this stage). Lessons learned during operational use systematically feed back into the Design and Testing stages, reinforcing the adaptive and iterative nature of the AI lifecycle. These continuous improvement loops enable sustained regulatory compliance, long-term system reliability, and continuous alignment with evolving trustworthiness requirements.

As each of the lifecycle stages have significant impact on the threat profiles, attack surfaces, and associated relevant Controls, it is important to understand how the various assets within the whole-system can be mapped to the primary lifecycle stage with which they are linked. It is likely the case that each asset within a whole-system assessed in-context would introduce distinct trustworthiness, security, or compliance risks at the different lifecycle stages. Understanding this ensures that risks and controls can be assessed consistently across the entire AI lifecycle.

The ENISA 2020 AI asset categorisation[119] (adopted in the FAITH AI TAF-v1 - Figure 7, p.92) suggests that the identification of assets within an AI ecosystem is a fundamental step in pinpointing what needs to be protected and what could possibly go wrong with an AI ecosystem. It defines the following high-level asset categories:

- Data
- Models
- Artifacts
- Environment / Tools
- Processes

However, the assets listed in the ENISA report are not linked to the AI lifecycle stage to which they are most relevant. This risks losing some of the threat and risk nuances that the lifecycle introduces. As a result, TAF-v2 has extended the ENISA list with additional Data and Model types described in the NIST AI RMF, including ones that are specifically associated with certain AI Model types, and then mapped all the assets to their most relevant lifecycle stage. This mapping enables lifecycle-specific risk identification, ensuring that threats and controls are assessed where they are most relevant.

Hence, Table 9 provides an indicative mapping between AI assets and the AI lifecycle stages. Assets may be relevant to multiple stages, but the mapping reflects their primary relevance. Additional items from the NIST AI RMF are indicated in blue. The Production stage encompasses runtime operation, monitoring, maintenance, performance evaluation, and end-of-life activities of the AI system.

[119] Baylon, C., Berghoff, C., Brunessaux, S., Burdalo, L., Dacquisto, G., Damiani, E., Herpig, S., Louveaux, C., Mistiaen, J., Polemi, N. and Praca, I., 2020. Artificial intelligence cybersecurity challenges; threat landscape for artificial intelligence.

Table 8: ENISA & NIST Asset-to-Lifecycle Mapping

Asset Category	Asset Example	Primary Lifecycle Stage
Data*	Raw Data	Design
	Training Dataset	Design / Testing
	Validation Dataset	Testing
	Testing Dataset	Testing
	Evaluation Dataset	Testing
	Operational / Inference Data	Production
	Pre-processed Data	Testing
	Synthetic Dataset	Testing
	Human Feedback Dataset	Production
	Metadata	Testing
	Pre-training Corpora (LLM-specific)	Design
	Instructional/Conversational Dataset (LLM-specific)	Design
	Retrieval Knowledgebase (RAG-specific)	Production
	Persistent Memory Stores (Conversational Agents)	Production
Models**	Trained Model	Testing
	Tuned Model	Testing
	Deployed Model	Production
	Foundational Model (LLMs)	Design
	Versioned Model	Design
	Agentic Model (Agentic AI-specific)	Production
	Safety / Policy Model (LLM-specific)	Production
Artifacts	Use Case Definition	Design
	Model Documentation	Testing
	Data Documentation	Testing
	Logs and Audit Trails	Production
	Model Performance Reports	Production
Environment / Tools	Data Exploration Tools	Design
	Training Platforms	Testing

	ML Libraries & Frameworks	Testing
	Deployment Infrastructure	Production
	Monitoring Tools	Production
Processes	Data Collection	Design
	Data Labeling	Testing
	Model Training	Testing
	Model Validation	Testing
	Model Deployment	Production
	Monitoring & Maintenance	Production
	Model Decommissioning	Production (End-of-life)
Actors / Stakeholders	Data Owner	Production
	AI Developer	Design
	Data Engineer	Testing
	End User	Production
	Data Provider	Testing
	Cloud Provider	Production
	Model Provider	Testing
	Service Consumer	Production

* It is also worth remembering that there are certain types of data (not AI-specific) that carry their own additional risks – in particular compliance risks as defined in the GDPR legislation. This includes Biometric data, Health data, Personal data, Sensitive data and Special Category data, all of which should also be individually considered when conducting risk assessment.

**NOTE: there is much more on model classification in section 3.5

In the context of the FAITH AI TAF-v2, the most relevant of the ENISA categories are Data, Models, Processes and Actors/Stakeholders (aka humans and/or organisations). For data, lifecycle stage matters because, for example, there may be different threats associated with a training dataset, such as systemic or representational bias that manifest during the Testing stage, compared with an output dataset used as the input for a second AI model, such as feedback loops, thresholding or binarization, that manifest during the Production stage. Equally, there are also differences in potential cascading threat vectors.

Similarly, for different model types the lifecycle stage is important, because, for example, an Agentic AI model may be subject to architectural design flaws at the Design stage that enable goal high-jacking or excessive agency further downstream, or it might display tool misuse resulting from hallucinations during the Production stage. In addition, each of these different threats have different levels of cascading effects too.

The same remains true for AI processes, which are the structured activities and operational workflows through which AI assets are created, transformed, deployed, monitored, and retired throughout the AI lifecycle. Examples of AI processes include data collection and preparation, data labeling, model training and validation, model deployment, performance monitoring, maintenance, and decommissioning. Processes are therefore dynamic in nature and constitute the mechanisms through which risks may be introduced, propagated, or mitigated. As before, different Processes at different lifecycle stages can also introduce different threats and provide different potential cascading threat vectors.

Finally, AI Actors/Stakeholders refers to the various individual, team and organisational level humans who interact with the AI system at the different lifecycle stages for different purposes. Each human (or team or organisation) has the potential to introduce different threats at different lifecycle stages. For example, developers can fail to sufficiently vet training data for bias, or make poor architectural choices, during the Design stage. Equally, Testing teams can underestimate drift or fail to sufficiently recreate real-world environments and stress scenarios. It is also possible that once an AI system is in production end users may attempt to misuse or jailbreak the system, or be socially engineered by adversaries. Clearly each of these provide opportunities for different potential cascading threat vectors at different points in the lifecycle. Indeed, humans in a system are frequently the most vulnerable of assets.

Understanding the relationship between assets and lifecycle stages enables the FAITH AI TAF-v2 to assess risks in a structured manner by identifying which assets are exposed, through which processes, and at which lifecycle stage. This approach supports precise threat modelling and facilitates the selection of appropriate technical and organisational Controls tailored to the lifecycle phase and asset type under assessment. Furthermore, the lifecycle-oriented approach adopted by the TAF-v2 recognises that the risk profiles associated with assets transition between lifecycle stages. For example, datasets initially used for model training may later be archived or replaced, while trained models may be updated, retrained, or retired. The inclusion of asset decommissioning and retirement within the lifecycle ensures that residual risks, such as data leakage or unintended reuse, are systematically addressed. By integrating asset classification and lifecycle mapping, a nuanced foundation for subsequent trustworthiness risk identification and trustworthiness assessment activities (that are consistent, explainable, and lifecycle-wide) can be provided.

4.1.1.5 Identify Trustworthiness Limitations

Many desired AI trustworthiness attributes cannot be guaranteed solely through technical controls because they arise from upstream structural conditions and systemic social practices, rather than from the AI system itself. In such cases, the best-designed control measures (e.g. reweighting, regularisation, explanation interfaces) can at most partially compensate for underlying injustices; they cannot fully eliminate the skew in the data-generating process or the power asymmetries that shape how AI is deployed and interpreted.

A health example illustrates this clearly. If minority groups systematically visit doctors less often, or present later and with more severe illness, then clinical datasets will underrepresent these groups, and may encode different disease trajectories and treatment pathways for them. This is not simply a data quality defect that can be patched by debiasing algorithms; it reflects long-standing barriers to healthcare access (financial, geographic, linguistic, cultural, or due to discrimination and mistrust). An AI model for diagnosis or triage trained on such data inherits these patterns even if the model architecture is fair-aware and direct sensitive attributes are removed. The introduction of Controls such as oversampling, group-fairness constraints, and post-hoc calibration can help mitigate this to some extent, but unless the underlying care pathways and participation patterns change, the AI system will remain constrained by what was never observed. In this sense, the trustworthiness attribute of Fairness, in particular Equity and Equality through demographic representativeness is structurally under-determined.

Similar dynamics arise in other domains. In employment, historical hiring decisions and performance evaluations may systematically favour majority groups or graduates from elite institutions. Even with rigorous feature vetting and fairness constraints, a hiring model trained on such records will reflect institutionalised preferences about who 'looks promising'. Control measures can reduce disparate error rates or remove some proxies, but they cannot on their own alter the fact that certain candidates never applied (because they self-selected out), were screened by human recruiters, or were never offered the kinds of roles that generate positive labels. As with the Health example, the Fair trustworthiness attribute, especially equal opportunity and non-discrimination, is a risk that cannot be fully mitigated through the application of social or technical Controls.

Further evidence of systemic Fairness risk can be found in AI models designed for the credit scoring and insurance industries. Here models rely on proxies for creditworthiness or risk calculations (e.g. income, employment history, postcode, prior claims) that are themselves shaped by housing segregation, educational inequality, and wealth gaps. Here, the aspiration that AI should not exacerbate financial exclusion cannot be satisfied merely by adding a fairness regulariser. As long as protected groups face systematically different structural constraints, any model optimising for default risk or expected cost will tend, unless explicit redistributive actions are implemented, to encode those constraints. Ensuring trustworthiness in terms of distributive justice and social inclusion may therefore require regulatory interventions (e.g. caps on the use of certain variables, universal service obligations, or subsidised lending), rather than only technical controls at model level.

Finally, in policing and criminal justice, predictive models trained on recorded crime or arrest data inherit the spatial and demographic biases of enforcement practices. Even if one applies state-of-the-art debiasing, the fact remains that police activity is often more intense in certain neighbourhoods than others, so crime is more frequently observed and documented there. As a result, trustworthiness attributes such as legitimacy and procedural fairness cannot be restored solely within the model; they are constrained by upstream policy choices about patrolling, surveillance, and criminalisation.

These examples underscore that certain dimensions of trustworthiness, particularly those related to Fairness and bias remain tied to structural representation, social power, and institutional practices As a result, they are only partially amenable to Control measures. Technical controls remain necessary, but they are insufficient. This wider understanding of the influence of societal inequalities forms an important part of providing a full socio-technical in-context risk assessment of an AI system. As a result, it is incumbent on a risk assessor to adequately recognise the impact of systemic and structural conditions on a Fairness assessment and to effectively communicate these equality and bias risks when reporting on AI system trustworthiness.

4.1.2 Phase 2: Threat Analysis

The second Phase of TAF-v2 involves the identification of threats. Within the FAITH project, the full range of Threats (and their associated Assets and Controls) are properly contained in the FAITH and SSM Knowlegdebases and as such will be fully reported in D2.4, where the knowledgebases will have been explicitly extended to include AI-specific threats, consequences and Controls.

Within the limitations of a framework however, the core principle, as expressed in TAF-v1, remains the identification of the frequency of a threat occurring, where frequency equates to the likelihood of that threat manifesting. The factors influencing the frequency of general cybersecurity threats occurring on individual assets in isolation, and the classification of the likelihood level in relation to the frequency (e.g. two attacks within a year = very high likelihood) with which it occurs is described in D2.1 section 4.2.2. However, TAF-v1 restricted itself to the threats that can manifest on a single asset in isolation.

In an extension to this, TAF-v2 suggests that identifying the likelihood (via a frequency judgement) for each of the threats associated with a whole-system in context - in other words, identifying the likelihood of that threat materialising in that specific context - is a way to accurately reflect the specific domain and context in which that asset exists, and thus enables an accurate in-context risk assessment. For example, the frequency (aka likelihood) of a data poisoning attack on a system deployed in a hospital context where security measures are already high due to the presence of Personal, Sensitive, Special category and general Health data is going to be lower than the likelihood of the same class of attack against a system in a small commercial business context. Hence, identifying context-specific likelihood is an important part of conducting an effective in-context risk assessment.

In a further extension to TAF-v1, TAF-v2 also introduces the need for risk assessors to identify the likelihood of AI-specific threats and the potential for cascading threats across the whole systems in the specific context in which it is embedded.

4.1.2.1 Identify Likelihood of Context-specific Threats

An important part of the in-context risk assessment process is to identify those threats which are most relevant to the specific context in which the AI system is located. This process involves

assessing the likelihood of those key threats occurring. TAF-v1 defined the threats to individual assets into two broad categories – technical threats and social threats – where social threats relate to the human assets and technical threats to the non-human assets. Of course, in reality the boundaries between the two are blurred, with some social threats (e.g. phishing) causing some technical threats (e.g. unauthorized system access or data poisoning), or wider societal issues (e.g. under-representation of minority groups in data) leading to technical threats (e.g. biased training data). Broadly however, the distinction between technical and social threats continues to serve the important role of focusing the mind of the risk assessor on not only those technical threats related to the model or data, but also to the people interacting with those assets and therefore remains relevant.

In TAF-v2, with its focus on in-context assessment, it is necessary to broaden the focus to include the likelihood of threats materializing within specific real-world contexts, which is an important and often underdeveloped component of whole-system risk assessment. Abstract threat catalogues and generic attack trees provide necessary baselines, but they systematically over- or under-estimate risk when divorced from concrete socio-technical conditions. Likelihood is not a property of a vulnerability in the abstract; it emerges from the interaction between technical exposure, attacker incentives and capabilities, organisational practices, and regulatory and sectoral environments.

In an AI ecosystem, this contextualisation is especially important because attack surfaces and incentives differ markedly across domains and architectures. For example, the practical likelihood of training-data poisoning, model inversion, or prompt injection depends on which interfaces are exposed, who can access them, how inputs are curated, and what value successful compromise would yield. The same vulnerability in a tightly controlled, offline medical imaging system and in a publicly accessible, API-driven foundation model will have radically different likelihood profiles. Failing to model these differences leads to misallocation of security resources and may create a false sense of trustworthiness.

Moreover, AI ecosystems typically involve heterogeneous actors—model providers, data holders, integrators, and end-user organisations—distributed across multiple jurisdictions. Each actor's security posture, incident response capability, and compliance culture shapes the probability that known threats will be exploited and successfully propagated. Context-specific assessment therefore requires mapping inter-organisational dependencies, human relationships, and governance responsibilities, and then estimating likelihood along these chains as well as at a single technical node.

Finally, regulatory and social context modulates attacker motivation and defensive investment trade-offs. High-value, safety-critical systems (e.g., healthcare, critical infrastructure, election-related systems) are more attractive targets and typically sit within stricter oversight regimes, altering both threat likelihood and detection probabilities.

A significant set of context-specific threats are centred around data assets and include threats related to data quality, integrity, and representativeness, with data poisoning, limited diversity in training datasets, lack of generalisability across operational settings, and concept or semantic drift being typical high-severity threats. However, the likelihood of these threats materializing is highly context-dependent. The same applies to regulatory, legal, and ethical non-compliance threats, which also constitute a major category of context-specific threats. These include discrimination and bias, improper data usage, consent violations, and privacy leakage, particularly when processing personal, sensitive, or video-based data. The likelihood of these threats becoming manifest is shaped by sector-specific legal obligations, data modalities, and societal expectations. While they may not always manifest as immediate technical failures, they can significantly undermine system legitimacy, user trust, and organisational accountability, especially in public-facing or safety-critical contexts. Equally, another prominent group of context-specific threats relates to limitations in interpretability, explainability, and transparency, and includes threats such as opaque model behaviour, insufficient rationale generation, or a lack of usable explanations for domain experts and operators. As before though, the likelihood of those threats becoming reality will vary from context to context. Finally, threats stemming from human-in-the-loop and organisational factors further shape the context-specific threat landscape. Lack of adequate training, unclear allocation of responsibilities, insufficient awareness of AI limitations, and weak human oversight mechanisms have been identified as high-impact threats, yet the likelihood that they will manifest will be significantly different in different contexts.

Overall, context-sensitive likelihood assessment of the most relevant threats thus becomes indispensable for grounding risk ratings, prioritising mitigations, and making defensible claims about AI trustworthiness.

4.1.2.2 Identify the Likelihood of AI-specific Threats

In section 3.6 above a range of AI-specific threats have been presented, including threats that are specific to LLMs and Agentic AIs and it is not the intention to repeat that here. However, it is important to emphasise that not only is the likelihood of the occurrence of general threats context-dependent, but so too is the likelihood of AI-specific threats. The context-specific nature of the AI threats will draw on the classification of the AI system undertaken in TAF-v2 Phase 1: Cartography, as the type of AI, the learning paradigms it employs, and importantly the function it serves within the specific context of use, will all affect the likelihood of threats materialising. There may also be, now or in the future, legislation, rules and regulations that frame the use of AI systems within a specific domain or context. It is therefore important for a risk assessor to identify the likelihoods of AI-specific threats as well as more general threats.

4.1.2.3 Identify the potential for Cascading threats

A broad discussion of the importance of cascading threats has been presented in section 3.4 above and will not be repeated here. However, as well as offering different threat likelihood

profiles, it is also the case that different contexts will present different opportunities for cascading threats across assets and beyond. Like with an assessment of the likelihood of AI-specific threats becoming manifest, so too can an assessment of the likelihood of cascading threats build on the classification of the AI system conducted in TAF-v2 Phase 1: Cartography. However in this case, it would be the classification of the interconnectedness of the system that matters. For example, AI systems with low interconnectedness (e.g. narrow AIs with no decision-making or acting capability run on an internal LAN) are significantly less likely to experience cascading effects than systems that are highly interconnected (e.g. an Agentic AI making and implementing operational decisions connected to the Web). Hence the likelihood of cascade effects also needs to be carefully considered.

4.1.3 Phase 3: Impact Assessment

The third Phase of TAF-v2 involves an assessment of impact that the key threats that have been identified in Phase 2: Threat Analysis would have on the specific context in which the AI system is located. TAF-v1 suggested that the impact, or consequence, of a threat materialising on an individual asset was important to consider in relation to the seven dimensions of trustworthinesss (especially the more technical dimensions) and presented a 6-point Impact Level rating to describe this (from Negligible to Very High). This remains an important part of the process, as it provides for the flow of information from the in-isolation asset-level assessment to the in-context whole-system assessment as the Impact Levels evaluated at that stage become the input into the whole system model, thus enabling model fine-tuning to specific contexts – thereby improving accuracy. TAF-v2, however, extends the Impact Assessment process to account for all assets in an AI ecosystem within the same risk assessment process.

4.1.3.1 Context-specific Impact Assessment

Identifying and assessing the full range of potential impacts (or consequences) of AI-related cybersecurity threats is a foundational step in any rigorous risk assessment, particularly in socio-technical environments where AI is tightly coupled to core business processes and social functions. Impacts should be conceived broadly to include technological, operational, business, legal, ethical, and reputational dimensions, as well as harm to individuals and wider society. These map closely to the full spectrum of trustworthiness dimensions. A narrow focus on purely technical fallout (for example, model unavailability or data corruption) can result in an underestimation of the true risk exposure created when AI system threats materialise.

However, that risk exposure will differ from one context to another, and from one AI ecosystem to another, For example, from a technological perspective, an AI cybersecurity incident can degrade model performance, distort training data, or disrupt pipelines, leading to silent failures rather than easily detectable outages. However, these consequences propagate into the business and operational space with differing levels of Impact. It is clear that a successful indirect prompt injection attack on a health AI system within a hospital context, causing a loss of reliability, carries a much higher level of Impact (e.g. potentially resulting in patient death) than the same attack on the customer complaints chatbot of an SME (e.g. loss of reputation), for example. This would

also hold true for any legal or compliance Impacts resulting from the same attack in those differing contexts.

Hence, systematically identifying context-specific impacts enables the risk assessor to calibrate severity levels that are meaningful for the domain in which the AI system operates and translate threat scenarios into concrete domain-specific consequences. In this way, a risk assessment can be fine-tuned to reflect an organisation's risk tolerance, the regulatory environment, and the vulnerability of affected populations. In turn, this contextualisation informs the prioritisation of Controls and monitoring, where critical risks are those whose consequences are unacceptable given the system's purpose and the profiles of those exposed to its decisions.

4.1.3.2 Cascading Threat Impact Assessment

Adopting a system-level view of Impact is crucial for trustworthy AI governance because AI systems rarely act in isolation. They are embedded in workflows, integrated into legacy IT, and mediated by human operators and institutional processes. Assessing impact only at the level of individual assets (for example, the model, a database, or a microservice) can obscure emergent harms that arise from interactions across the system. The compromise of an apparently low-value component (e.g. a logging subsystem or a prompt interface) may cascade through dependencies and lead to more significant Impact and problematic misbehaviours at the level of organisational decisions or public-facing services. In other words, as Threats cascade through a system, so too can Impact become magnified or multiplied. System-level impact assessment should capture these propagation pathways by assigning Impact Levels to each of the assets in the system that reflect the specific context and then using these to calculate an overall Impact Level for that Threats path, thus giving an accurate Impact Level assessment for that Threat in that system in that context.

This system perspective, that accounts for cascading Threats and multiplying Impact, is also more effective for making trustworthiness decisions, because trustworthiness is ultimately judged at the level of the AI Ecosystem level and its real-world effects, not only at the level of individual technical artefacts. Stakeholders care whether the overall socio-technical system behaves reliably, fairly, and safely under credible threat scenarios. By modelling Impact across the full system, the assessor can evaluate whether residual risks after mitigation are compatible with the organisation's commitments on robustness, fairness, accountability, and respect for rights. Asset-level assessment remains necessary as a building block, but only system-level Impact reasoning provides the substrate for defensible, context-sensitive judgements about whether an AI system can be considered trustworthy in its actual context.

4.1.4 Phase 4: Vulnerability Analysis

The fourth Phase of TAF-v2 involves an assessment of all the AI-specific Vulnerabilities and the associated Controls and Control Strategies that can be implemented to protect against those Vulnerabilities. This builds on the two previous Phases in the TAF: Phase 2: Threat Analysis and Phase 3: Impact Assessment. TAF-v1 suggests that Vulnerability Levels are context-dependent.

This remains the case. It also presented a method for assessing the overall Vulnerability Level, also using the 6-point rating system, based on the total percentage of possible Controls implemented at the asset-level. TAF-v2, however, adds more detail by suggesting the need for the risk assessor to consider Controls related to technical Vulnerabilities and to human/social Vulnerabilities as interrelated at a whole-system level, thus recognising cross-cutting Controls and the notion of Control Strategies - a suite of individual Controls that address Vulnerabilities at the system-level. TAF-v2 also introduces the need for trade-off decision-making at this stage when considering Controls to implement in particular.

4.1.4.1 Identify Technical Vulnerabilities and Controls

It is important that all AI-specific Technical Vulnerabilities and Controls are identified. This involves their systematic identification, with the objective not only to catalogue them, but also to identify which technical Controls are currently already in place in the system, and to then assess whether those Controls are sufficient and effective within the operational and organisational context of the AI system and align with the organisation's risk tolerance.

Common technical vulnerabilities in AI systems cluster around data integrity, model robustness, system interfaces, and the AI interconnectedness with other systems, each with associated Controls that must be understood in a holistic, system-wide manner. Depending on the AI Lifecycle stage, key vulnerability classes include failures, weaknesses, or limitations in model design, training and other data issues, input manipulation, data and model poisoning, model extraction and inference attacks, and insecure integration of AI components into broader architectures. There are considerable on-going efforts by many organisations (e.g. the OWASP GenAI Security Project[120], ENISA[121], NIST AI RMF) to identify and classify the AI-specific Vulnerabilities and Controls, hence it is important for risk assessors to remain up-to-date with the current state-of-the-art mitigation strategies, including preventive, detective, and corrective measures applied at different stages of the AI lifecycle.

Furthermore, it is also the case that different technical Controls (or lack thereof) will impact different trustworthiness dimensions. For example, a failure to implement Controls for Data Poisoning such as Provenance Tracking measures (see **Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε.**) would negatively impact the Validity & Reliability trustworthiness attribute. As a result, each technical Vulnerability and Control should be understood in a holistic, system-wide manner as contributing to system-level Control Strategies that aim to improve overall trustworthiness from a technical perspective.

Table 10 summarises some example technical vulnerabilities, the AI Lifecycle stage to which they apply, their associated Controls, and some of the limitations those Controls have. Considering Control limitations supports a more realistic Vulnerability assessment by supporting the trade-

[120] https://genai.owasp.org/

[121] Baylon, C., Berghoff, C., Brunessaux, S., Burdalo, L., D'Acquisto, G., Damiani, E., Herpig, S., Louveaux, C., Mistiaen, J., Polemi, N. and Praca, I., 2021. Securing machine learning algorithms.

off decision-making process. It is fully recognized that some Vulnerabilities can relate to multiple AI Lifecycle stages, but here an assessment is presented of the most relevant stage.

Table 9: Example AI-specific Technical Vulnerabilities, Controls and Limitations across AI Lifecycle Stages

Technical vulnerability	AI Lifecycle Stage	Identified Controls	Control Limitations
Insufficient consideration of evasion attacks	Design	Robust model design; defensive architectures	Design-time measures may not protect against inference-time attacks
Use of widely known or standard model architectures	Design	Secure development and hardening practices	Does not prevent model extraction or transfer attacks
Data poisoning during training or fine-tuning	Testing	Curated datasets; data provenance tracking; anomaly detection	Provenance does not prevent poisoning at trusted sources
Absence of adversarial training	Testing	Adversarial training techniques	Typically limited to known attack classes; risk of performance degradation
Opaque decision or optimisation logic	Testing	Logging of scoring and optimisation functions	Logging alone does not guarantee explainability or accountability
Low robustness to real-world data variability	Testing	Synthetic data generation; staged testing	Synthetic data may not reflect real operational complexity
Lack of documentation of model assumptions	Testing	Model cards; lifecycle documentation	Documentation can become outdated or inconsistently applied
Ineffective explanation methods for end users	Testing	Usability testing of explanations	Episodic testing may not reflect real-world usage
Bias from skewed or incomplete training data	Testing	Bias audits; fairness testing	Requires sensitive attributes; audits are typically periodic

Inability to assess fairness due to anonymisation	Testing	Privacy-preserving fairness techniques	Persistent trade-offs between privacy and fairness objectives
Limited generalisability to unseen environments	Production	Diverse and multi-source datasets	High cost; rare or extreme conditions may remain uncovered
Insufficient model drift monitoring and control	Production	Performance monitoring; drift detection	Detection may occur after degradation has already impacted outcomes
Reduced interpretability of model outputs	Production	Saliency maps; contextual metadata	Outputs may be unusable or misleading for non-expert users
Lack of robust isolation between system prompts and untrusted user or content prompts in LLM pipelines, enabling prompt injection	Production	Prompt separation; fine-tuning against override attempts	Fragile against indirect or multi-step prompt attacks
Prompt leakage	Production	Prompt rotation; probing detection	Detection is probabilistic and can be bypassed through slow probing
Unsafe or unconstrained action generation	Production	Rule-based constraints; safety maps	Static constraints may fail in dynamic environments
Lack of data minimisation, masking, and retention controls for sensitive data in AI model training data, telemetry, and logs	Production	Encryption; access control; anonymisation	Does not fully mitigate inference or reconstruction attacks

Dependence on external or cloud computation	Production	Anonymisation; privacy-preserving transfer	Exposure to network and infrastructure-level attacks remains
Over-trust in external data sources that are injected into the LLM context without intent-checking or isolation	Production	Source labelling; restriction of agent autonomy	Requires consistent enforcement across ingestion pipelines
Over-reliance on unverified generative outputs without human-in-the-loop or retrieval-augmented verification	Production	Retrieval-augmented generation; entropy-based checks; human-in-the-loop	Dependence on external knowledge sources; increased system complexity
Lack of detection of abnormal or adversarial inputs	Production	Input validation; anomaly detection tools; sanitization; conservative confidence thresholds	Often reactive; limited ability to detect novel or adaptive adversarial patterns
Susceptibility of AI model inference endpoints to model extraction and inference attacks	Production	Rate limiting; access controls; perturbation of outputs	Too-high limit thresholds; compromised accounts; repeated or adaptive querying

*NOTE: these and other AI-specific Vulnerabilities and Controls can be found in the extended FAITH and SSM Knowlegdebases that will be reported in D2.4

While not exhaustive, what Table 10 shows is the interdependencies between technical Vulnerabilities and Controls. Understanding these vulnerabilities only at the level of individual assets is not wholly sufficient, because their exploitation depends on how components interact across the full AI system. For example, adversarial input validation at the API boundary can be undermined if poisoned data enters via a separate ETL pipeline; controls on model or data

artefacts are ineffective if the pipelines by which they connect to other software assets are insecure; and rate-limiting model queries does not meaningfully reduce extraction risk if multiple microservices expose the same model through different routes. This interdependence means that Controls should be identified and implemented as a coordinated package - as Control Strategies - not just as isolated technical fixes.

4.1.4.2 Identify Human Vulnerabilities and Controls

Having identified the technical Vulnerabilities and Controls associated with an AI Ecosystem, it is also vital that all Human and/or Social Controls are also identified. This involves the systematic identification of human-related Vulnerabilities and their associated Controls, with the objective not only to catalogue them, but also to identify which human/social Controls are currently already in place in the system, and to then assess whether current Controls are sufficient and effective within the operational and organisational context of the AI system and align with the organisation's risk tolerance.

Human/social Controls address Vulnerabilities that relate to human behaviour, organisational practices, and socio-technical interactions. This recognizes the fact that in many cases trust failures in AI systems originate in human decision-making, coordination gaps, or organisational immaturity rather than in solely in technical model considerations. Equally, human/social Vulnerabilities and Controls may also be role-sensitive. Developers, operators, decision-makers, and governance actors are exposed to different Vulnerabilities and therefore may require tailored Controls in order to remain effective in complex socio-technical settings and avoid a one-size-fits-all approach.

Common human/social Vulnerabilities in the human assets in AI ecosystems cluster around three broad classes: Cognitive; Behavioural; Organisational. The first of these manifest at the individual level, the second at the teams level, and the third at the organisational level. As with technical Vulnerabilities and Controls, there is a degree of overlap between these three classes. Equally, the Controls associated with each class Vulnerability may also connect with other human/social Controls and/or have Technical solutions as well as human ones. It is also the case that different Controls (or lack thereof) will impact different trustworthiness dimensions. For example, a failure to implement Controls for Automation Bias such as Explainability measures (see **Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε.**) would negatively impact both the Explainability & Interpretabiliy and Safety trustworthiness attributes. As a result, each human/social Vulnerability and Control should be understood in a holistic, system-wide manner as contributing to system-level Control Strategies that aim to improve overall trustworthiness from a human/social perspective.

Table 11 summarises some of the key human/social Vulnerabilities, the AI Lifecycle stage to which they apply, their associated Controls, and some of the limitations those Controls have. Considering Control limitations supports a more realistic Vulnerability assessment by supporting the trade-off decision-making process. It is fully recognized that some Vulnerabilities can relate to multiple AI Lifecycle stages, but here an assessment is presented of the most relevant stage.

Table 10: AI-specific Human/Social Vulnerabilities, Controls and Limitations

Vulnerability Class	Human / Social vulnerability	AI Lifecycle Stage	Identified Controls	Control Limitations
Cognitive	Lack of knowledge, awareness or understanding of AI-specific threats	Production	security and AI-literacy training; phishing and social-engineering simulations; clear AI use policies	Training decay over time; habituation; new AI attack techniques (e.g., AI-driven phishing)
	Automation bias / over-reliance on AI outputs	Production	Human-in-the-loop review; decision support not decision replacement; calibrated uncertainty displays; explainability structures	Users may still over-trust confident-sounding systems; explanations can be misinterpreted or ignored; process controls bypassed under time pressure or workload
	Confirmation bias	Testing	Diverse review teams; red-teaming and adversarial testing; structured hypothesis testing; independent audits	Organisational culture and incentives may still reward "good news"; teams may discount or rationalise conflicting signals; and red-teaming/audits are periodic rather than continuous
	Social-engineering susceptibility and trust miscalculation	Production	multi-factor authentication; behavioural anomaly detection; UI cues that distinguish verified from unverified sources; "zero-trust" access patterns	Habituation to MFA prompts and UI cues; time and workload pressures
	Overconfidence and complacency	Production	Regular independent risk assessments, external penetration tests, purple-team exercises (collaborative cybersecurity simulations)	Overconfidence can extend to the assessments themselves; executives may deprioritise uncomfortable findings; security

				trade-offs miscalculated
	Non-compliance with organisational AI and security policies	Production	Clear, concise AI and security policies that specify permitted tools, data classes, and use cases, supported by leadership communication; technical enforcement methods	Formal policies may be poorly understood or perceived as misaligned with day-to-day work, encouraging quiet non-compliance; Imperfect or overly rigid technical controls pushing users towards unsupervised channels
Behavioural	Informal practices and workarounds	Design	Standardised templates for AI workflows, prompt libraries, and documentation; DevEx tools that support formal practices; peer code review	Templates and tools may constrain innovation and creativity and promote generic one-size-fits-all solutions; Reviews become box-ticking exercises
	Risk-taking and over-experimentation with AI tools	Testing	Sandboxed experimentation environments with synthetic or de-identified data; Guardrails in prompts and orchestration layers; Peer review and change-management processes	Guardrails can be circumvented by creative users making them hard to test; guardrails can erode as systems become more complex and integrated which is also hard to test; under-pressure teams may quietly deploy experimental patterns without full sandboxing
	Unclear accountability and responsibility gaps	All lifecycle stages	Responsible, Accountable, Consulted, Informed (RACI) roles specific to AI workflows; AI governance boards with cross-functional membership;	RACI becomes outdated as teams/roles change; lack of governance board

			escalation paths and post-incident review processes	enforcement power over day-to-day decisions; crisis-pressure overrides formal processes
Organisational	Insufficient segregation of duties	Testing	Mandatory peer/cross-team reviews at key milestones; separation of environments (dev/test/sandbox) with distinct access controls; independent reporting outside development chains	Resource constraints force teams to adopt multiple roles
	Misaligned incentives and priorities	All lifecycle stages	Balanced KPIs incorporating security and trustworthiness alongside performance; explicit executive focus on linking AI trustworthiness to organisational risk tolerance and long-term reputation	Short-term business pressures consistently override long-term risk considerations
	Limited oversight capacity	All lifecycle stages	Graduated oversight processes; automated risk assessment; central governance boards with cross-functional membership; 3rd Party audit and certification	Budget competition with core business priorities; requires executive commitment and integration into business KPIs
	Lack of documentation and knowledge transfer	All lifecycle stages, particularly handovers between stages	Documentation-as-code practices; mandatory 'AI handoff checklists'; monitoring for undocumented aspects	Documentation decays as priorities shift; checklists become ritualised and incomplete; tacit knowledge resists formal capture

By implementing appropriate human/social Controls it is possible reduce the likelihood that human Vulnerabilities will trigger or increase cascading failures across interconnected AI systems. They also improve the organisation's ability to detect early warning signals, intervene appropriately, and recover from incidents in a transparent and controlled manner.

The outcome of this phase is a structured understanding of human/social Vulnerabilities and their associated Controls at the human asset level (Individual / Team / Organisation), which can be carried forward into a system-level risk assessment. This ensures that the AI system is correctly treated as a socio-technical system in which human/social Vulnerabilities are treated with the

same analytical rigour as technical Vulnerabilities with human/social Controls becoming part of system-wide Control Strategies.

4.1.4.3 Identify Control Strategies

A Control Strategy constitutes a deliberate, integrated portfolio of preventive, detective, and/or corrective measures designed to manage identified risks to an acceptable level, with particular emphasis on addressing the unique threat landscape of data dependencies, model fragility, dynamic inputs/outputs, and socio-technical whole-system interactions. Unlike isolated controls, a Control Strategy coordinates complementary mechanisms across system layers to create defence-in-depth, ensuring that the compromise of any single point does not cascade into systemic failure.

In AI contexts, Control Strategies would ideally span the full lifecycle and bundle technical, procedural, human, and organisational Controls together to effectively treat threats like prompt injection, data poisoning, and model extraction holistically. Consider the threat of prompt injection targeting a customer-facing chatbot integrated with internal APIs. Ensuring effective security training for designers and developers (Control 1 [human]: Training) will lead to the implementation of a standalone rate limit on the chatbot endpoint (Control 2 [technical]: API gateway throttling) that reduces query volume. However, this still allows slow, persistent attacks, which is picked up as a result of mandatory peer/cross-team reviews (Control 3 [human]: oversight & segregation) during testing. Bundling in input sanitisation and validation (Control 4 [technical]: pre-model content filters and schema enforcement) blocks malicious payloads, while least-privilege tool access (Control 5 [technical]: sandboxed API calls with scoped tokens) contains any successful injection. Adding runtime behavioural monitoring (Control 6 [technical]: anomaly detection) and output perturbation (Control 7 [technical]: response rounding and filtering) provides layered containment. Finally, incident response automation (Control 8 [technical]: isolation of compromised sessions) ensures rapid recovery, and escalation paths and post incident review processes (Control 9 [human]: accountability and responsibility) ensures lessons are learned and change is introduced if necessary.

This bundling across assets is primarily intended to prevent this example Threat from propagating across assets and systems. A successful injection at the interface is neutered by privilege limits before reaching sensitive APIs; anomalous tool use triggers isolation before data exfiltration; and automated playbooks limit manual error in response. The strategy's effectiveness derives from Controls at different AI Lifecycle stages, interlocking technical coverage, and the integration of human/social Controls into the mix.

The power of such strategies lies in minimising single points of failure and maximising compensating controls. As detailed in sections 4.1.4.1 and 4.1.4.2, no individual Control is perfect, however, through the identification and implementation of Control Strategies attackers are forced to chain multiple exploits if they wish to be successful, thus dramatically increasing the technical knowledge and skill required to undertake them. This is particularly critical for AI systems, where subtle compromises (e.g., backdoored weights) can propagate silently through inference chains, downstream decisions, and human trust calibration until manifesting as the

types of Very High Impact failures that have been identified in TAF-v2 Phase 3: Impact Assessment as particularly relevant for the context in which the AI ecosystem is situated. Ultimately, Control Strategies for AI elevate security from tactical point fixes to strategic system resilience, enabling trustworthy operation in adversarial environments.

4.1.5 Phase 5: Risk Analysis

The fifth Phase of TAF-v2 involves unifying and operationalising the previous Phases into a layered risk assessment activity, including ensuring the flow of information across layers, specifically from in-isolation assessment to in-context assessment. TAF-v1 proposed the use of a 6-point risk assessment scale (Negligible to Very High) in order to align with earlier Phase ratings, and introduced an overall risk calculator process, which remains valid. TAF-v2 extends this with the introduction of the integration of information from the in-isolation assessment layer to the in-context assessment layer, and the linking of FAITH tools that supports the operationalizing of the two layers.

4.1.5.1 Asset-level In-isolation Risk Assessment

At this stage of the FAITH AI TAF-v2, the risk assessor conducts an in-isolation risk assessment of core AI assets, focusing in particular on data and model assets as previously identified in the Cartography phase (Section 4.1.1). Importantly, this includes conducting an assessment of the of the team maturity of the team owner of each asset, as previously described in D2.1 / TAF-v1. The objective of this step is to evaluate the intrinsic risk profile of each asset independently from its operational environment, system integrations, or organisational processes. By isolating individual assets, the assessor can determine the inherent vulnerabilities, threat exposure, and trustworthiness characteristics attributable specifically to each AI asset itself.

There is a range of existing AI risk assessment guidelines, methodologies and standards that can help structure such an in-isolation risk assessment – many of which have been explored and integrated into the two FAITH TAF versions, including in this document. There is also a range of asset-level risk assessment tools available, but for tools that are aligned with the FAITH TAF-v1 and v2 it is recommended that risk assessors use the tools developed and validated within the FAITH project. Specifically, this includes the AI Model Hub (and its associated AI Model and Data Passports) for describing and documenting model and data assets; trustGUARD for risk assessing the trustworthiness of model and data assets and receiving recommended asset-level Controls; and trustSENSE for Human Actor Profiling, Team Maturity Assessment, human-based risk scoring, and recommended human/social Controls. These technical outputs of the FAITH project compliment the framework developed and presented here but will be discussed in detail in the D2.4 Technical Report.

Importantly, the in-isolation risk assessment process using TAF-compliant FAITH tools produces information that can be used to inform the in-context risk assessment in such a way that it becomes fine-tuned for the precise context and conditions in which the AI ecosystem is situated. For example, Phase 1: Cartography can usefully inform the right assets to model in the whole-

system (including model assets such as Jurisdictions, AI Model type, and the humans who interact with the system). Next, Phase 2: Threat Analysis supports the fine-tuning of Likelihood Levels for those key in-isolation assets (models and datasets) in the system model. Likewise, Phase 3: Impact Assessment enables the fine-tuning of the Impact Levels for those same assets. Then, Phase 4: Vulnerability Assessment allows those existing Controls at the asset-level to be accurately modelled at the system level. Together, this information flow from the in-isolation assessment layer to the in-context assessment improves analytical precision and reduces the likelihood of conflating component vulnerabilities with systemic amplification effects. This layered methodology ultimately strengthens the overall reliability, transparency, and accountability of the AI risk assessment process and supports nuanced and accurate trustworthiness assessment.

4.1.5.2 Modelling Assets In Context for Whole-system Risk Assessment

Conducting risk assessments for AI trustworthiness next requires a holistic system model that captures not only the technical components but also the socio-technical ecosystem in which the AI operates. This whole-system modelling process delineates all assets within the wider system, including data, processes, software, hardware, human actors, physical and digital locations, and jurisdictions, in order to identify risks that transcend isolated model evaluation, especially risks related to cascading threats. Unlike narrow audits focused on algorithms or datasets, whole-system in-context modelling reveals how trustworthiness attributes are co-determined by upstream data practices, deployment contexts, and downstream interpretations. It unifies the previous TAF Phases. As a result, when conducting a whole-system in-context risk assessment it is important to model all these assets effectively. A range of AI-specific assets have been presented in section 4.1.1.4 above and this provides a framework that can guide a risk assessor when undertaking system modelling. In addition, the classification of the Model and the identification of the humans in the system conducted in Phase 1: Cartography can be used to establish assets to include in the system model.

However, it is not sufficient to merely identify the assets in an AI ecosystem. Just as important are the exact relationships each asset has to another, as the nature of these relationships have a heavy influence on determining the threats and cascading threats paths through the system. For example, an AI Model will ‘read’ some data as input or training data (which may contain biases), while it may also ‘create’ some data as output (which may amplify those biases). This opens the way for biases in training data to propagate outwards to human and non-human decision-making, thus reducing the Fairness trustworthiness attribute. Equally, data may be ‘stored’ in a database and may ‘relate to’ some humans. The database may be compromised and Personal, Sensitive, Biometric or any other data may be poisoned. The poisoned data may then in turn be ‘used by’ an AI model, leading to a negative impact on the Reliability trustworthiness attribute. Similarly, the Robustness trustworthiness attributes is impacted when Processes are ‘hosted’ on servers which are ‘connected to’ the internet, thereby opening attack vectors that would not be present if the server were only connected to a local network, or where humans ‘manage’ or ‘operate’ devices that they use to ‘interact with’ data, thus opening the potential for social engineering attacks that gain system or device access and introduce malicious behaviours.

Furthermore, it is also not enough to just model the assets and their relationships without taking into consideration the specific context in which the AI model is situated. As defined in the ISO27000(x) series of standards[122], risk is calculated as the likelihood of an event occurring and the impact that event would have if it were to occur. In particular, whole-system modelling needs to account for the impact of a threat becoming a reality within the specific context in which the assessment is being carried out. This will involve the risk assessor using the output of the TAF-v2 Phase 2: Threat Analysis and Phase 3: Impact Analysis to accurately adjust the Likelihood and Impact Levels for different threats so as to accurately reflect their specific context. This process of fine-tuning the system model can be further enhanced by using the results of the in-isolation risk assessments on the Model, Data and Human assets to adjust the risk levels associated with specific trustworthiness attributes for those assets in the system model.

Finally, whole-system modelling also requires the risk assessor to include any existing Controls that are present in the system, identified in Phase 4: Vulnerability Assessment, as the presence or not of those Controls will significantly affect the risk levels across assets in the system. Again, it may be the case that in-isolation risk assessments have recommended the introduction of Controls directly at the asset level which, once implemented, need to be modelled at the system level. However, it may also be that Controls are more removed from an asset than that yet are still intended to protect that individual asset. For example, securing a server room with strict biometric access (Control 1) for a restricted whitelist of authorised humans only (Control 2) does not involve the server itself but, clearly, are designed to (indirectly) protect the server (not the room or the humans). These too need to be modelled.

The importance of whole-system in-context modelling cannot be overstated as it enables all the relevant assets, relationships, threats / cascading threats, and Controls to be identified and Likelihood and Impact Levels to be set that accurately reflect specific contexts. It also unifies the information flow from in-isolation risk assessment to in-context risk assessment. Hence, the results of a whole-system in-context risk assessment provide accurate, nuanced, system-wide and asset-level information that enables the identification of the primary threats and the most effective Controls to mitigate them. This empowers decision-makers to make the necessary trade-offs to leave their system with an acceptable level of residual risk and with a sufficiently high level of trustworthiness that is appropriate to their context.

4.1.5.3 System-level In-context Risk Assessment

At this stage of the FAITH AI TAF-v2, the risk assessor conducts an In-context risk assessment of all the assets modelled as a connected ecosystem. The objective of this step is to evaluate the total risk profile of the whole AI ecosystem in order to reasonably account for cascading primary and secondary Threats and escalating Impact, as well as recommending Control Strategies that

[122] https://www.iso.org/standard/iso-iec-27000-family

address system-wide Vulnerabilities - with the ultimate output being a trustworthiness assessment for the whole system across all trustworthiness dimensions.

Although there are some system-level risk assessment tools, such as ThreatModeler[123] or Microsoft's Threat Modeling Tool[124], none:

- align with the TAF-v2
- are able to model the full range of assets (software, hardware, processes, network infrastructure, humans, jurisdictions…etc) and all their relationships
- are able to account for assets at all lifecycle stages
- allow fine-tuning of all the Likelihood and Impact Levels associated with the trustworthiness attributes of all assets to properly reflect deep context (and domain) customisation

As a result, the FAITH project has supported the further development and extension of just such a tool - the System Security Modeller (aka Spyderisk), which enables an automated risk assessment to be conducted that properly accounts for cascading primary and secondary Threats and recommends Control Strategies (and Controls) that are capable of treating a Threat across multiple linked assets in the system.

Ultimately, by integrating the outputs of the in-isolation step as fine-tuning inputs of the in-context step, this serves as the final operationalised risk assessment activity that accounts for human/social and technical, context-customised, system-wide Threats, Impacts and recommended Controls that provides the structured evidence base required for effective trade-off discussions and the implementation of the most effective mitigations that properly align with organisational risk tolerances.

4.1.6 Phase 6: Countermeasures

Once the in-isolation and in-context risk assessments have been completed, the sixth and final Phase of TAF-v2 involves using the risk assessment outputs as a decision-making evidence base with varying levels of granularity, ranging from a top level trustworthiness assessment across the seven attributes, through system-level detailed risk assessment scores, trustworthiness assessments and Control Strategies, to asset-level granular risk assessment scores, trustworthiness assessments and Controls. At this point it is possible to have fully informed risk tolerance and mitigation trade-off discussions at senior level, including involving non-expert stakeholders.

4.1.6.1 Determine Trade-offs and Risk Tolerance

Section 3.7 has presented the way in which there are trade-offs at different levels when considering approaches to maximise the trustworthiness of an AI ecosystem. This includes trade-

[123] https://threatmodeler.com/
[124] https://learn.microsoft.com/en-us/azure/security/develop/threat-modeling-tool

offs between trustworthiness attributes, as well as trade-offs between trustworthiness, operational capabilities, performance requirements, effort and financial costs, and knock-on effects to linked systems or organisations.

Asset-level and system-level risk assessment outputs provide complementary insights essential for calibrating organisational risk tolerance and determining Control implementation, that requires deliberate synthesis to maximise decision-making accuracy. Asset-level analysis yields granular prioritisation of Control implementation on critical components. However, it optimises Control implementation decisions to align these with the Control Strategies recommended by system-level assessments that mitigate emergent propagation risks, interdependencies, and cascading consequences through mitigations that affect multiple assets simultaneously. Key considerations include reconciling micro-level severity with macro-level impact. For example, a low-value asset may warrant minimal Controls in isolation, yet demand harder mitigation actions if positioned as an asset from which a primary Threat may originate.

Equally, risk tolerance calibration proceeds through using the two levels of risk assessment outputs to better inform scenario mapping, where Very-High-Risk Threats can be translated into concrete business outcomes (e.g. revenue loss, regulatory fines, reputational harm) and where quantitative thresholds (e.g., maximum tolerable downtime, acceptable error rates under attack) and qualitative tolerances (e.g., reputational risk tolerance) will further guide Control Strategy bundling.

The process entails:

1. the aggregation of residual risks post-controls
2. stress-testing against worst-case cascades
3. stakeholder alignment via governance forums
4. dynamic review as tolerance evolves with business context

This ensures controls are neither over-engineered (wasting resources) nor under-provisioned (exposing the organisation), achieving cost-effective trustworthiness aligned with strategic risk limits.

4.1.6.2 Implement Controls and Control Strategies

This stage of the TAF-v2 simply involves the implementation of the decisions reached during the trade-off discussions. Clearly, each specific context will result in a bespoke Control implementation plan complete with timings.

4.1.6.3 Re-run Risk Assessments

The final activity suggested by TAF-v2 is that once one or more of the Controls have been implemented the Risk Analysis process (Phase 5) should be re-run to provide an up-to-date in-isolation and/or in-context risk assessment in order to evidence how the introduction of those Controls have impacted the trustworthiness of the AI ecosystem. Ideally this should be done using the same risk assessment tools. Both the trustGUARD tool for asset-level risk assessment and the System Security Modeller for system=level assessment, not only allow on-going post-

control re-assessment, but also enable risk assessors to model/account for the presence of a Control before it is actually implemented in the real. This means these tools can be used by trade-off decision-making teams to analyse the impact on trustworthiness of multiple possible Control implementation plans (at asset and system level) before actually undertaking the work.

5 Implementation and Validation of the FAITH AI TAF-v2

5.1 TAF-v2 Validation Plan

The objective of the FAITH AI TAF-v2 validation process is to robustly evaluate its consistency, usability, and practical utility (usefulness) for different stakeholders. This is a validation of the TAF-v2 only, not of the FAITH tools that implement the TAF-v2, which will be reported on in D2.4. The key stakeholder groups will be:

1. A representative from the AI developer teams associated with each LSP
2. The named risk assessor from each LSP
3. A user representative (non-technical) from each LSP

The methodology will involve three online 60-minute workshops (one for each stakeholder group). The first part of the workshop will provide a walk-through of the TAF-v2, the second will present structured qualitative and quantitative feedback opportunities. Feedback opportunities will be structured on adapted Technology Acceptance Model criteria and consist of an online quantitative survey and qualitative oral feedback. The validation process will be implemented from M30 to M36.

The workshop schedule and content generation will adequately consider challenges including:

- time zone differences
- language and communication clarity - including the potential for minor but critical misunderstanding of key terms
- cultural and/or demographic norms - especially to ensure equal contribution
- technical infrastructure - such as access to feedback surveys
- potential power imbalances
- facilitator bias - including in the creation of content

5.2 TAF-v2 Implementation in FAITH Tools

One of the most critical aspects that TAF-v2 has introduced is the information flow between the in-isolation risk assessment (and associated tools) that were presented in TAF-v1 and the in-context risk assessment (and associated tool) presented here. In this regard, TAF-v2 is firstly being implemented through the integration of trustGUARD and the System Security Modeller (SSM) tools. Although this is not explicitly detailed in the Grant Agreement as a Task, nevertheless, the WP2 team believe that it is a vital aspect of supporting a nuanced, contextual customisation and an important part of implementing TAF-v2. Consequently, processes are being developed by which SSM can automatically ingest the relevant information captured in trustGUARD concerning the Impact Levels associated with individual data and model assets according to trustworthiness attribute to automatically fine-tune the Impact Levels for the corresponding asset in the system model. Equally, SSM will also ingest information about the existence of asset-level Controls captured by trustGUARD and automatically mark those Controls as present in the system. This

involves some complex mapping between the two knowlegdebases that underpin them, as well as some considerable technical integration and development activity to enable (and act on) the data flows. More will be reported on this in D2.4.

In addition, as formal feedback from the LSPs is captured during their use of the in-isolation tools during the first pilot scenarios, TAF-v2 will be used to align and interpret that feedback so as to inform the iterative development of those tools ready for use during the in-context assessment phase within the second (replication) pilot scenarios.

6 Conclusions

The FAITH AI Trustworthiness Assessment Framework version 2 constitutes a significant advance from version 1. Through a process of collaboration, TAF-v2 has been enhanced to improve its applicability to real-world AI systems through targeted extensions. These extensions include:

- A move from the assessment of isolated assets (data, model, human) to that of ecosystems of connected assets (other software and processes, hardware and network infrastructure, people, places and jurisdictions)
- An explicit accounting of cascading and propagating risks across systems
- The inclusion of context-specific fine-tuning of Impact Levels and existing Controls for all assets in a system model that allows for deep context-customisation - this is particularly enabled for key assets that have been the subject of in-isolation risk assessment (data, models, humans) by the integration of information flows from in-isolation to in-context risk assessments
- A more detailed understanding and classification of AI Model types and their still-emerging associated Threats and Controls
- The inclusion of system-level Control Strategies
- A move from single-dimension trustworthiness optimisation to balanced, evidence-based trade-off management across trustworthiness attributes and operational constraints
- Refined and redefined steps for each of the 6 TAF Phases - along with the need to document actions taken at each step (see Figure 1)

Despite this expanded scope, the foundational principles of the framework remain consistent with those introduced in TAF-v1. FAITH AI TAF-v2 retains a human-centric approach, continues to align with established risk management and AI governance standards (in particular the principles articulated by ENISA, NIST and ISO, noting any updates and evolutions of them), maintains the foundational Layer 1 assumptions detailed in TAF-v1 (except for the exclusion of cascading threats), and retains the three AI Lifecycle stages, the 6-point risk assessment ratings (negligible to very high) and the 6 top-level TAF Phases.

As a result, the extensions and developments presented in TAF v2 offer a more holistic and operationally applicable framework for assessing trustworthiness in complex real-world AI systems, while preserving fundamental methodological continuity with the first version.

TAF-v2 will be formally validated with three groups of stakeholders (developers, risk assessors and non-technical users) and incorporated into the technical iterative development process for the second versions of the FAITH Tools that will be used in the second pilot phase.