Ref. Ares(2026)2930006 - 18/03/2026

# D3.3

## Validation Report

| | |
|---|---|
| **Related Work Package** | **WP3 – Communication, Dissemination, Exploitation & Impact** |
| **Related Task** | Task 3.3. - Evaluation of Methodology (NIST) & Policy Recommendations |
| **Lead Beneficiary** | FORTH |
| **Contributing Beneficiaries** | All partners involved in Task 3.3 |
| **Document version** | v.1.0 |
| **Deliverable Type** | R-Document, Report |
| **Distribution level** | PU-Public |
| **Contractual Date of Delivery** | 30/11/2025 |
| **Actual Date of Delivery** | 18/03/2026 |



| | |
|---|---|
| **Authors** | Dimitrios I. Fotiadis, Manolis Tsiknakis, Kostantinos Marias, Grigoris Kalliatakis, Nikolaos S. Tachos, Vasileios C. Pezoulas (FORTH)<br>Nic Fair (UOS)<br>Sara Colantonio, Maria Antonietta Pascali, Danila Germanese, Claudia Caudai (CNR) |
| **Contributors** | Eleni Tsalapati (ATC), Giorgos Giotis (ATC) |
| **Reviewers** | Evangelia Anagnostopoulou (ICCS), Alberto Topini (UNIFI) |



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 |
|---|---|---|---|
| **v0.1** | TOC circulation | 03/10/2025 | FORTH |
| **v0.2** | Initial structure and methodology drafting | 20/11/2025 | UoS |
| **v0.3** | NIST AI RMF mapping section drafted | 10/01/2026 | UoS |
| **v0.4** | FAITH tools validation methodology drafted | 25/01/2026 | FORTH |
| **v0.5** | Integration of partner inputs | 10/02/2026 | FORTH,CNR UoS |
| **v0.6** | Integration of partner inputs | 18/02/2026 | FORTH,CNR UoS |
| **v0.7** | Integration of partner inputs | 25/02/2026 | ATC, CNR |
| **v0.8** | Consortium review version circulated | 05/03/2026 | FORTH |
| **v0.9** | Final revisions and formatting | 16/03/2026 | UOS |
| **v1.0** | Final version submitted | 18/03/2026 | FORTH |

## Statement of Originality

This deliverable contains original unpublished work except where clearly indicated otherwise.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

## Disclaimer

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

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

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

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

## Executive summary

This deliverable presents the Validation Report of the FAITH project, documenting the methodology and activities undertaken to evaluate the FAITH ecosystem in alignment with the NIST Artificial Intelligence Risk Management Framework (AI RMF), with particular focus on the MEASURE function. The objective of this report is to demonstrate how the FAITH tools, frameworks, and methodologies support systematic testing, evaluation, verification, and validation (TEVV) of AI systems while contributing to trustworthy and responsible AI deployment across diverse real-world domains.

The validation approach adopted in FAITH reflects the project's multi-layered structure. FAITH develops a Trustworthiness Assessment Framework (TAF), a set of AI risk assessment tools, and deploys these solutions in seven Large Scale Pilots (LSPs) across heterogeneous domains. Consequently, validation activities operate at multiple levels: (i) validation of the conceptual framework and Knowledgebase underpinning the FAITH ecosystem, (ii) validation of the FAITH tools and services against the NIST AI RMF MEASURE requirements, and (iii) validation through their application in real-world pilot environments.

A core contribution of this deliverable is the systematic mapping between the NIST AI RMF MEASURE function and the FAITH Trustworthiness Attributes (TWAs) and tools. The MEASURE function emphasises the need to transform AI risk hypotheses into empirical evidence through quantitative and qualitative evaluation methods. FAITH operationalises this principle by providing tools that enable structured risk assessment across key dimensions of AI trustworthiness. Through this mapping, the deliverable demonstrates how FAITH supports eleven out of the thirteen MEASURE sub-functions, providing mechanisms for documentation, human-centred evaluation, context-aware performance testing, security and safety risk assessment, fairness and bias evaluation, and privacy compliance analysis.

The report also introduces the FAITH Tools Validation Methodology, which follows a three-tier validation approach. The first tier focuses on validating the foundational elements of the ecosystem, namely the FAITH TAF and the FAITH Knowledgebase. This includes a structured co-design process with LSP stakeholders, enabling the enrichment of the Knowledgebase with domain-specific threats, vulnerabilities, and mitigation controls derived from real deployment scenarios. The second tier validates the FAITH tools against the NIST AI RMF MEASURE framework to ensure methodological alignment with internationally recognised risk management standards. The third tier implements a mixed-method evaluation strategy, combining technical testing, pilot deployment, and structured feedback collection from stakeholders.

Within this framework, FAITH tools such as TRUSTGUARD, trustSENSE, SSM, the AI Model Passport, the Data Passport, and the FAITH AI Model Hub play complementary roles in enabling risk assessment both in isolation and in context. Pre-deployment dry-run testing, pilot-specific implementation plans, and structured feedback mechanisms ensure that the tools are evaluated across diverse operational environments and that user feedback contributes to iterative improvement.

Finally, the validation results feed into the development of policy recommendations, translating empirical findings from the pilots into insights relevant for regulators, standardisation bodies, and organisations deploying AI systems. These recommendations highlight the importance of socio-technical evaluation, participatory risk assessment, and standards-aligned governance mechanisms to support the responsible adoption of trustworthy AI across Europe.

Overall, this deliverable demonstrates that the FAITH ecosystem provides a standards-aligned, evidence-based approach for assessing and managing AI risks, bridging the gap between conceptual AI governance frameworks and their practical application in real-world systems.

# Table of Contents

1. Introduction ..... 9
1.1 Objectives and Scope ..... 9
1.2 Structure of the document ..... 9
2. FAITH – NIST Validation Alignment ..... 11
2.1 Understanding the NIST AI RMF – MEASURE Function ..... 11
2.2 Mapping the NIST AI RMF MEASURES to FAITH ..... 11
2.2.1 Measure 2.1: Test sets, metrics, and details about the tools used during test, evaluation, validation, and verification (TEVV) are documented ..... 13
2.2.2 Measure 2.2: Evaluations involving human subjects meet applicable requirements and are representative of the relevant population ..... 14
2.2.3 Measure 2.3: AI system performance or assurance criteria are measured qualitatively or quantitatively and demonstrated for deployment setting(s) ..... 15
2.2.4 Measure 2.4: The functionality and behavior of the AI system and its components are monitored when in production ..... 15
2.2.5 Measure 2.5: The AI system to be deployed is demonstrated to be valid and reliable ..... 16
2.2.6 Measure 2.6: AI system is evaluated regularly for safety risks ..... 17
2.2.7 Measure 2.7: AI system security and resilience are evaluated ..... 17
2.2.8 Measure 2.8: Risks associated with transparency and are examined ..... 18
2.2.9 Measure 2.9: The AI model is explained, validated, and documented, and AI system output is interpreted within its context ..... 19
2.2.10 Measure 2.10: Privacy risk of the AI system is examined ..... 19
2.2.11 Measure 2.11: Fairness and bias are evaluated ..... 20
2.2.12 Measure 2.12: Environmental impact and sustainability of AI model training and management activities are assessed ..... 21
2.2.13 Measure 2.13: Effectiveness of the employed TEVV metrics and processes in the MEASURE function are evaluated ..... 21
2.3 Summary ..... 22
3. FAITH Tools Validation Methodology ..... 24
3.1 Overview of FAITH Tools Validation Approach ..... 24
3.2 Validation of the FAITH AI Trustworthiness Assessment Framework and the FAITH Knowledgebase ..... 24
3.3 Validation of the FAITH Tools Against NIST AI RMF MEASURE Function ..... 26
3.4 Mixed-methods FAITH Tool Validation Plan ..... 26
3.4.1 Dry Run Testing ..... 27
3.4.2 FAITH Tool Validation Plan with LSPs ..... 27
3.4.3 Formal Feedback and Iteration ..... 27

4. Policy Recommendations ..... 30
4.1 Turning Feedback Into Insight ..... 30
5. Conclusions and Next Steps ..... 33
5.1 Overall validation conclusions ..... 33
5.2 Recommendations for future evaluations ..... 34
5.3 Link to subsequent deliverables and exploitation activities ..... 35
Appendix A ..... 37
Section 1: About You ..... 37
Section 2: Perceived Usefulness ..... 38
Section 3: Ease Of Use ..... 39
Section 4: Human Factors & Social / Organisation Influences ..... 40
Section 5: Support for NIST MEASURES ..... 41
Section 6: Detailed Suggestions ..... 42

## List of Tables

Table 1. Mapping between NIST AI RMF MEASURE sub-functions (2.1–2.13) and the FAITH Trustworthiness Attributes (TWAs) together with the FAITH tools/services that support their evaluation. N/A indicates that the FAITH ecosystem does not directly support the corresponding measure, while None denotes that no specific FAITH tool addresses it. ..... 22

## List of Figures

Figure 1: The 4 Core NIST AI RMF Functions. ..... 12
Figure 2: FAITH Validation Approach Overview. ..... 24
Figure 3: The KB co-design journey with the LSPs ..... 26

## List of Abbreviations

| Abbreviation | Explanation |
|---|---|
| **AI** | Artificial Intelligence |
| **AI RMF** | Artificial Intelligence Risk Management Framework |
| **AI SUT** | Artificial Intelligence System Under Test |
| **DBDM** | Data Bias Detection and Mitigation |
| **ENISA** | European Union Agency for Cybersecurity |
| **EU** | European Union |
| **GDPR** | General Data Protection Regulation |
| **HCI** | Human–Computer Interaction |
| **KB** | Knowledgebase |
| **LLM** | Large Language Model |
| **LSP** | Large Scale Pilot |
| **MDR** | Medical Device Regulation |
| **NIST** | National Institute of Standards and Technology |
| **PETs** | Privacy Enhancing Technologies |
| **RMF** | Risk Management Framework |
| **SSM** | System Security Modeller |
| **TAM** | Technology Acceptance Model |
| **TAF** | Trustworthiness Assessment Framework |
| **TEVV** | Test, Evaluation, Verification and Validation |
| **TWA** | Trustworthiness Attribute |
| **UI** | User Interface |
| **UX** | User Experience |
| **WP** | Work Package |

# 1. Introduction

## 1.1 Objectives and Scope

The aim of this deliverable is to report on the development of the validation methodology for the FAITH project in line with the NIST Artificial Intelligence Risk Management Framework (AI RMF)[1] and to indicate how this may be used to inform Policy Recommendations. It reports on the activity in T3.3 (WP3), which is scheduled to run from M26 to M42 and hence this is an active Task and therefore subject to future evolution.

Initially, it is important to note that the FAITH project has multiple parallel activity strands – firstly, it is developing the FAITH AI Trustworthiness Assessment Framework (TAF); secondly it is implementing and iteratively developing a suite of AI risk assessment tools, many based on an advanced risk Knowledgebase (FAITH KB); thirdly it has multiple real-world use cases (Large Scale Pilots – LSPs) across seven different domains where one or more AI systems are being deployed and used in either sandbox or real-world situations.

As a result, validation takes on a variety of meanings within the FAITH project at a number of levels. It can refer to the validation of the FAITH TAF – a theoretical framework requiring certain, mainly qualitative, validation activities. This corresponds to the first tier, and it is reported mainly in WP2 D2.2, as explained later on in this deliverable.

Nonetheless, validation also refers to the validation of the FAITH tools themselves – both demonstrating how they support NIST AI RMF MEASURE function validation actions for AI systems, and how they can themselves be validated. This will form the main focus of this deliverable.

Furthermore, validation may also refer to the validation of the deployed AI systems within the LSPs. This may involve the application of the NIST AI RMF MEASURE function validation actions through the use of the FAITH tools in many cases. This deliverable will report on the plans and activities for the validation of FAITH tools in the LSPs, which will form a secondary focus for this deliverable.

## 1.2 Structure of the document

This deliverable is structured to provide a coherent and progressively layered account of the FAITH validation methodology, its alignment with the NIST AI Risk Management Framework (AI RMF) MEASURE function, and the translation of evaluation findings into actionable policy recommendations.

Following this introductory chapter, which defines the objectives, scope, and contextual positioning of validation activities within the FAITH project, **Chapter 2** presents a detailed analytical mapping between the FAITH ecosystem and the NIST AI RMF MEASURE function. It begins with an overview of the conceptual foundations of the MEASURE function and its role within the broader NIST AI RMF lifecycle (GOVERN–MAP–MEASURE–MANAGE). It then systematically examines each relevant MEASURE sub-function and demonstrates how FAITH Trustworthiness Attributes (TWAs), tools, and services support or operationalise these requirements. This mapping establishes the standards-aligned basis of the FAITH validation approach and clarifies the extent to which FAITH enables evidence-based AI risk measurement across technical and socio-technical dimensions.

**Chapter 3** describes the FAITH Tools Validation Methodology itself. It adopts a multi-tier validation perspective and distinguishes between: validation of the foundational FAITH AI Trustworthiness Assessment Framework (TAF) and Knowledgebase (KB), validation of FAITH tools against the NIST AI RMF MEASURE function, and mixed-method validation through structured deployment in Large-Scale Pilot (LSP) environments. This chapter details the co-design journey undertaken with LSP stakeholders

[1] https://airc.nist.gov/airmf-resources/playbook/

to validate and enrich the FAITH Knowledgebase, the dry-run testing phase conducted prior to tool deployment, the structured implementation plans developed with each pilot, and the formalised feedback and iteration mechanism grounded in established evaluation models such as the Technology Acceptance Model (TAM). By documenting both technical validation and user-centred evaluation processes, Chapter 3 demonstrates that validation within FAITH extends beyond methodological compliance to include usability, contextual adaptability, and organisational acceptance.

**Chapter 4** builds upon the empirical findings of the validation activities to derive policy recommendations. It explains how structured feedback, cross-domain risk patterns, and implementation challenges are synthesised into evidence-based recommendations aimed at European governance bodies and standardisation actors. The chapter emphasises the socio-technical character of AI trustworthiness and articulates how lessons learned from the FAITH Large-Scale Pilots can inform future regulatory alignment, harmonisation efforts, and operational guidance for trustworthy AI deployment.

Finally, **Chapter 5** presents the overall validation conclusions and outlines next steps. It summarises the degree of alignment achieved with the NIST AI RMF MEASURE function, reflects on the strengths and limitations of the current validation scope, and proposes directions for future evaluation cycles and potential extensions of the FAITH ecosystem.

# 2. FAITH – NIST Validation Alignment

## 2.1 Understanding the NIST AI RMF – MEASURE Function

The NIST Artificial Intelligence Risk Management Framework (AI RMF)[2] provides four core functions, of which the MEASURE function is the third. The NIST AI RMF suggests that firstly there is a need for the development and adoption of organisation-wide AI policies and roles (the GOVERN function). Next, it is vital to characterise the specific AI system under test, including its task/purpose and its lifecycle; identify the stakeholders and impacted communities; discover potential benefits, harms and context-specific risks; and decide whether the AI system is justified when compared with non-AI alternatives (the MAP function). Thereafter, once the AI system is chosen for deployment, who, what and how to test that system becomes important (the MEASURE function), before finally decisions are taken about which risks should be prioritised and how they should be mitigated (the MANAGE function).

In particular, the MEASURE function grew out of NIST's long-running work on the testing and evaluation of complex systems and was shaped during the multi-year development of the AI Risk Management Framework (AI RMF) following a U.S. congressional mandate in the 2020 National Artificial Intelligence Initiative Act. NIST adopts the principle that it is not possible to effectively manage what you do not measure, and that rigorous test, evaluation, verification, and validation (TEVV) are central to risk management. Early AI standards work and "zero-drafts" around TEVV for AI systems reinforced the need for a dedicated function in the AI RMF that would focus on measurement and evidence, not just governance principles. Validation under the MEASURE function is essentially about turning risk hypotheses into evidence by designing, running, and interpreting tests that show how the AI behaves against the trustworthiness characteristics defined in the NIST AI RMF MEASURE function.

As NIST developed the MEASURE function through stakeholder engagement gaps in systematic AI evaluation, particularly around social impacts, fairness, robustness, and emergent behaviours were identified - resulting in NIST making socio-technical effects, not only technical performance, primary objects of measurement. The result is that one of the four core functions—MEASURE— explicitly includes both quantitative and qualitative assessment. MEASURE emphasises that tests must reflect the system's actual context and affected populations, not be only a convenient benchmark. NIST explicitly encourages involving users and impacted groups in evaluating how AI systems work in practice. Methods include surveys, interviews, focus groups, contextual inquiry, diary studies, and observational studies of real or simulated use. Findings are typically coded into themes (e.g. "over-reliance on system advice", "confusing explanation language") and linked back to specific risks in MEASURE categories.

It should be noted that the NIST AI RMF MEASURE function is intended to be applied to an actual AI system-under-test (AI SUT), however, within the FAITH project it is necessary to demonstrate how the NIST MEASURE function validation actions are effectively supported by the suite of FAITH tools and services, rather than directly applying them to the AI SUTs deployed in the Large Scale Pilots (LSPs).

## 2.2 Mapping the NIST AI RMF MEASURES to FAITH

THE FAITH AI Trustworthiness Assessment Framework (TAF) identifies 3 main phases in the lifecycle of an AI system – design / testing / production, and 6 Phases in the framework for assessing trustworthiness – Cartography / Threat Analysis / Impact Assessment / Vulnerability Analysis / Risk Assessment / Countermeasures. Similarly, the NIST AI RMF also breaks down the process of risk

[2] https://airc.nist.gov/airmf-resources/playbook/

assessment across the lifecycle of an AI system into four core functions that loosely parallel the FAITH TAF (see Figure 1):

- GOVERN – Develop organisational policies for AI risk management
- MAP - Characterise the system / tasks / lifecycle stage; Identify stakeholders / benefits / harms / context-specific risks
- MEASURE – Socio-technical, mixed-method testing - what, who, how to measure AI risk
- MANAGE – Decide which risks to prioritise and how to respond




*Figure 1: The 4 Core NIST AI RMF Functions.*

The NIST AI RMF provides the foundation for systematically assessing, analyzing, and tracking AI risks throughout the lifecycle, operationalised by establishing rigorous practices for testing, evaluation, verification, and validation (TEVV), supported by appropriate metrics, representative data, and continuous assurance mechanisms.

While the risk assessment tools developed within the FAITH project allow risk assessment at each of three different AI lifecycle phases, when the FAITH tools are used to support validation activities as described in this Chapter, they are positioned at the deployment lifecycle stage, the Risk Assessment TAF Phase and the MEASURE function – MEASURE 2 of the NIST AI RMF.

The NIST AI RMF groups the MEASURE function into four broad categories, each with multiple sub-measures:

1. **Measure 1: Identify and apply appropriate metrics and methods**
    a. Measure 1.1: Approaches and metrics for measurement of AI risks enumerated during the MAP function are selected for implementation
    b. Measure 1.2: Appropriateness of AI metrics and effectiveness of existing controls is regularly assessed and updated
    c. Measure 1.3: External stakeholders and communities are involved in regular assessments as necessary per organisational risk tolerance
2. **Measure 2: AI systems are evaluated for trustworthy characteristics** – this part of the MEASURE function is of particular importance to the FAITH project as will be expanded on in the following sub-sections below.
3. **Measure 3: Mechanisms for tracking identified AI risks over time are in place**
    a. Measure 3.1: Approaches, personnel, and documentation are in place to regularly identify and track existing, unanticipated, and emergent AI risks in deployed contexts

   b. Measure 3.2: Risk tracking approaches are considered for settings where AI risks are difficult to assess using currently available measurement techniques or where metrics are not yet available
   c. Measure 3.3: Feedback processes for end users and impacted communities to report problems and appeal system outcomes are established and integrated into AI system evaluation metrics
4. **Measure 4: Feedback about efficacy of measurement is gathered and assessed**
   a. Measure 4.1: Measurement approaches for identifying AI risks are connected to deployment context(s)
   b. Measure 4.2: Measurement results regarding AI system trustworthiness in deployment context(s) and across AI lifecycle are informed by input from domain experts and other relevant AI actors
   c. Measure 4.3: Measurable performance improvements or declines based on consultations with relevant AI actors including affected communities, and field data about context-relevant risks and trustworthiness characteristics, are identified

Together, these Categories/Measures specify how organisations should turn AI-risk hypotheses into empirical and experiential evidence using both quantitative and qualitative assessments. These Measures address performance, reliability, generalisability, safety, security, privacy, and in-production monitoring, while embedding transparency, accountability, explainability, and fairness assessments. Collectively, they enable evidence-based measurement of AI trustworthiness and risk, supporting informed deployment decisions and ongoing risk mitigation in line with the AI RMF.

The Measure 2 Category is of particular importance to the FAITH project because the project aims to develop, deploy and evaluate a set of tools designed to identify trustworthiness characteristics of an AI SUT through risk assessment. Therefore, it is important that the 13 sub-measures within MEASURE 2 are effectively mapped to the 7 Trustworthiness Attributes (TWAs) that have been identified in the FAITH Trustworthiness Assessment Framework (FAITH AI TAF - see D2.1) and to the FAITH risk assessment tools and services themselves – thereby demonstrating how the FAITH project supports the NIST MEASURE function. This will now be detailed in the following sub-sections:

### 2.2.1 Measure 2.1: Test sets, metrics, and details about the tools used during test, evaluation, validation, and verification (TEVV) are documented

This involves the systematic documentation of the measurement approaches, test sets, metrics...etc that have been identified as appropriate in NIST AI RMF MEASURE 1. This provides a foundation for a valid, reliable, repeatable measurement process and contributes to ensuring transparency, reproducibility, and scientific rigour in AI system evaluation. Detailed recording of measurement procedures facilitates comparison across models, versions, and deployment contexts, which is vital for lifecycle management and continuous improvement. Moreover, such systematic documentation underpins regulatory compliance and ethical accountability, as it provides auditable evidence that validation has been conducted according to predefined, appropriate standards. As a result, all subsequent MEASURE 2 sub-measures explicitly include reference to the need to document all activities. This maps to two FAITH TWAs – Validity & Reliability and Transparency & Accountability.

Key **suggested actions** here include following industry best practice for documentation of testing; regularly assessing the effectiveness of tools used for documentation purposes; and updating the tools as required. FAITH has implemented these suggested validation actions across **the AI Model Passport**, and **the Data Passport**, under the aegis of **the FAITH AI Model Hub**, which collectively serve as the authoritative repository for TEVV-related evidence. These tools consolidate information on datasets, model versions, training and validation procedures, applied metrics, and observed limitations, linking them directly to the corresponding trustworthiness risks and controls. In addition, the **TrustGuard** and

**SSM** risk assessment tools also provide version histories of each risk assessment and system model, allowing the on-going and permanent documentation of risk assessments and results. In addition, the FAITH **Risk Report Compiler**, which will be developed later in the project will provide users with a consolidated report of the risks, impacts and controls identified in the outputs of the TrustGuard and SSM tools as documented evidence of risk assessment outcomes during the validation process. As a result, within FAITH, TEVV documentation is not treated as a static reporting activity but as living artifacts that evolve throughout the model lifecycle of design, development, deployment, and monitoring. This approach supports consistency with the NIST MEASURE 2.1 function by enabling continuous validation of AI systems' reliability and supporting informed decision-making for risk management, mitigation actions, and cross-domain comparability within the FAITH ecosystem.

| NIST MEASURE | FAITH Trustworthiness Attribute (TWA) | FAITH Tool or Service |
|---|---|---|
| Measure 2.1: Documentation | Validity & Reliability<br>Transparency & Accountability | AI Model Passport<br>Data Passport<br>FAITH AI Model Hub<br>TrustGuard<br>SSM<br>Risk Report Compiler |

### 2.2.2 Measure 2.2: Evaluations involving human subjects meet applicable requirements and are representative of the relevant population

This involves the protection of human participants in testing activities according to domain, jurisdictional and legal requirements, as well as the need to have representative human participants and/or test data about humans. Measurement and evaluation of AI systems often involve direct interaction with human subjects or the use of data derived from human participants. In such cases, appropriate human subject protection measures are required, including safeguarding participants' welfare and interests, minimizing risks, and ensuring compliance with mandatory legal and ethical requirements. In parallel, validation actions should reflect the characteristics of the populations affected by or interacting with the AI system, as non-representative data or subject groups may lead to inaccurate assessments or harmful outcomes. This maps to the two FAITH TWAs – Safety and Fairness.

Key **suggested actions** here emphasize responsible human-subject involvement and robust representativeness analysis throughout data collection and evaluation. This includes following informed consent and privacy norms and standards; using various methods to ensure representativeness (discrepancy analysis / disaggregation / thresholds and alerts…etc); and fair dataset creation, IP, privacy, quality and bias-checking processes. In alignment with this, FAITH adopts a **human-centric approach** to the deployment of the AI systems-under-test (AI SUT) across the 7 LSPs, ensuring that all relevant ethical, legal, and organisational requirements are met. Human-in-the-loop evaluations and the use of human-derived data are designed to be representative of the intended user populations and operational contexts, accounting for diversity in roles, expertise, demographic characteristics, and interaction modalities. Where required when sensitive or personal data is being used by the AI SUT, the SUT is deployed in a sandbox environment (e.g. in the Healthcare LSP).

FAITH has further implemented these suggested validation actions through the **trustSENSE** tool which is specifically designed to enable an assessment of the maturity of the core stakeholders – developers and users – in relation to their trustworthiness when designing/developing and/or using AI systems. In this regard, trustSENSE uses a set of Maturity Dimensions to capture human-centric trust and safety dimensions, including: (i) Proactivity & Threat Awareness, (ii) Responsibility & Ethics, (iii) Innovation &

Adaptability, (iv) Resilience, (v) Collaboration & Knowledge Sharing, (vi) Technical Proficiency, and (vii) Policy Adherence, across representative stakeholder groups. By integrating trustSENSE outcomes with human-centric AI SUT deployment approaches, FAITH supports consistency with the NIST MEASURE 2.2 function requirements for representative, safe, and compliant human-centered AI validation.

| NIST MEASURE | FAITH Trustworthiness Attribute (TWA) | FAITH Tool or Service |
|---|---|---|
| Measure 2.2: Human Safety and Representation | Safety<br>Fairness | trustSENSE |

### 2.2.3 Measure 2.3: AI system performance or assurance criteria are measured qualitatively or quantitatively and demonstrated for deployment setting(s)

This involves the deeper understanding of deployment context of use in order to directly evaluate risks and impacts in real world environments thus providing a more accurate understanding of how and under what conditions deployed AI creates impacts, who is most likely to be impacted, and what that experience is like. Robust validation demands a rich understanding of the specific deployment context of use, including tasks, users, organisational processes, and socio-technical dependencies. Only by grounding evaluation in these situated conditions can one directly assess how an AI system behaves under realistic constraints, data distributions, and interaction patterns, and thereby estimate risks and impacts with fidelity. This Measure maps to the FAITH TWA – Validity & Reliability.

Key **suggested actions** here include using quantitative metrics, such as error rates, calibration, training times, prediction latency, subgroup performance differentials, or AI actor competency and experience assessment to provide structured evidence on system reliability and potential disparities across populations. Furthermore, using complementary qualitative methods, including expert review, sustained community engagement, UI/UX testing, and analyses of lived experience of diverse and representative stakeholders can enable an effective in-context assessment of the AI SUT.

FAITH operationalises this concept through the **SSM** tool which is specifically designed to enable the entire system to be modelled. This includes modelling the people, places, processes, data, hardware, and network infrastructure that accurately reflects the specific context of the deployed AI SUT. For each asset in the system model, it is also possible to adjust the impact level that a particular threat to that asset would have in that precise real-world deployment context were that threat to materialise. These adjustments to impact levels to reflect reality are informed by the quantitative risk assessment conducted in **TrustGuard** and the qualitative team maturity assessment conducted in **trustSENSE**, and by integrating the outputs of the FAITH assessment tools in this way, FAITH supports consistency with the NIST MEASURE 2.3 function requirements for mixed-method AI system testing that accurately reflect real-world deployment contexts.

| NIST MEASURE | FAITH Trustworthiness Attribute (TWA) | FAITH Tool or Service |
|---|---|---|
| Measure 2.3: Mixed-method in-context testing | Validation & Reliability | SSM |

### 2.2.4 Measure 2.4: The functionality and behavior of the AI system and its components are monitored when in production

This involves the monitoring of new issues (e.g. drift), risks or emergent behaviours (e.g. personas) over time once an AI system is deployed. Continuous monitoring of the functionality and behaviour of artificial intelligence (AI) systems after deployment helps to ensure their reliability, safety, and ethical compliance in real-world environments. While pre-deployment testing can identify many performance

issues, the inherent complexity and adaptivity of AI models mean that unforeseen behaviours often emerge in dynamic operational contexts. Monitoring enables early detection of performance drift, bias amplification, or unintended interactions with users and other systems. Furthermore, systematic observation supports accountability and regulatory compliance by generating traceable evidence of system decisions and actions. Aspects of this map onto three FAITH TWAs – Validation & Reliability; Safety; Accountability & Transparency – in particular Reliability and Accountability.

Key **suggested actions** here include pre- and post-deployment performance comparison; hypothesis testing; anomaly detection; alerting; output testing; and crucially human oversight and review in order to minimise error propagation and feedback loop risks. Although the FAITH **trustSENSE**, **TrustGuard** and **SSM** tools can all be used at multiple points in the AI Lifecycle, including on multiple single occasions during the deployment phase (e.g. after a new control is introduced or another AI model or new software or process is introduced into the system), none of them are designed to support real-time performance or behaviour monitoring. As a result, the FAITH project does not support the real-time monitoring of AI performance and behaviour risks.

| NIST MEASURE | FAITH Trustworthiness Attribute (TWA) | FAITH Tool or Service |
|---|---|---|
| Measure 2.4: The functionality and behavior of the AI system and its components are monitored when in production | Validation & Reliability<br>Safety<br>Accountability & Transparency | N/A |

### 2.2.5 Measure 2.5: The AI system to be deployed is demonstrated to be valid and reliable

This involves the careful use of validation indicators, especially where proxy indicators are used for phenomena that cannot be directly measured, that have the highest likelihood of measuring the properties or impacts under assessment. Two factors are important, firstly that the indicators do not result in spurious correlations or fail to account for error propagation and cascading impacts on connected systems; and secondly that sufficient consideration is given to exploring the potential for the AI system to be used or to develop behaviours for which it was not originally designed. This maps directly to the FAITH TWA – Validity & Reliability.

Key **suggested actions** here include establishing the AI system's operational conditions, socio-technical context and limits; reviewing and assessing the appropriacy of validation activities (is it measuring what it claims to, is it influenced by other factors, is it generalisable beyond pre-deployment training/testing, is it statistically / experimentally robust…etc); identifying robustness and reliability indicators; conformity with industry standards; flexibility beyond defined limits; and human oversight. FAITH operationalises aspects of this concept through the **SSM** tool where operational conditions and socio-technical context of the AI SUT can be modelled as an interconnected network where the influence and/or cascading effect of one network asset on another can be risk assessed. This risk assessment is underpinned by the **FAITH and (closely aligned) SSM Knowledgebases** where a range of AI-specific threats, including, importantly, cascading threats and specific threats arising from linked AI (and IT) systems, have been modelled. In this way, FAITH supports consistency with aspects of the NIST MEASURE 2.5 function requirements, namely for paying attention to cascading effects within operational conditions and socio-technical contexts.

| NIST MEASURE | FAITH Trustworthiness Attribute (TWA) | FAITH Tool or Service |
|---|---|---|

| | | |
|---|---|---|
| Measure 2.5: The AI system to be deployed is demonstrated to be valid and reliable | Validation & Reliability | SSM |

### 2.2.6 Measure 2.6: AI system is evaluated regularly for safety risks

This involves the regular demonstration of the safety of an AI system, including ensuring that residual negative risk continues to not exceed risk tolerance levels and that the system is reliable and robust. Ensuring safety involves the systematic integration of repeatable quantitative risk analysis, formal verification techniques, and on-going and updating control and mitigation strategies, the documentation of which is capable of evidencing that the system operates within acceptable risk boundaries. From a governance and trust perspective, transparent documentation of risk assessment supports on-going regulatory compliance and ethical accountability, particularly when AI decisions affect human welfare or public infrastructure. This maps directly to the FAITH TWA – Safety

Key **suggested actions** here include repeatedly measuring activities for safety in development and deployment contexts, including stress-testing, chaos testing, risk assessment; developing incident response plans; collecting and sharing relevant statistics and evidence; and using the NIST AI RMF MANAGE function to guide risk management actions. FAITH operationalises aspects of this concept through the **TrustGuard** and **SSM** tools, both of which enable highly repeatable risk assessment at both the development and deployment phases. Both tools are also capable of demonstrating the change to risk profiles and the associated acceptable risk boundaries once mitigation controls have been introduced, thereby enabling the on-going assessment of safety. In addition, the FAITH **Risk Report Compiler**, which will be developed later in the project will provide users with a consolidated report of the risks, impacts and controls identified in the outputs of the TrustGuard and SSM tools. In this way, FAITH supports consistency with the NIST MEASURE 2.6 function requirements to document and assess the safety of an AI system from a risk perspective on an on-going basis in both development and deployment phases.

| NIST MEASURE | FAITH Trustworthiness Attribute (TWA) | FAITH Tool or Service |
|---|---|---|
| Measure 2.6: AI system is evaluated regularly for safety risks | Safety | TrustGuard<br>SSM<br>Risk Report Compiler |

### 2.2.7 Measure 2.7: AI system security and resilience are evaluated

This involves assessing whether the AI system and the ecosystem within which it is deployed is capable of maintaining their functions and behaviours when confronted with unexpected, changed or adverse events. Such assessment goes beyond conventional performance testing by examining how the system responds to distributional shifts, component failures, or emergent interactions within socio-technical environments. Importantly, resilience assessment involves identifying potential cascading failure modes. Furthermore, evaluating resilience at both system and ecosystem levels supports the design of appropriate safeguarding mechanisms, including fallback procedures, graceful degradation strategies, and human-in-the-loop interventions. This is particularly important in safety-critical and high-stakes domains, where loss of functionality may lead to significant harm, legal liability, or erosion of public confidence. This maps directly to the FAITH TWA – Security & Resilience.

Key **suggested actions** here include implementing security tests, including red-team exercises, failure mode assessment and cybersecurity threat risk assessment; identifying and implementing countermeasures (i.e. security controls and control strategies that mitigate security risks); verifying that connected 3rd party AI resources and personnel have also undertaken robust cybersecurity risk assessment and mitigation actions; and sharing experience and learning with others within and

without the organisation. FAITH directly operationalises this Measure through the **TrustGuard** and **SSM** tools which are primarily designed to provide a comprehensive cybersecurity risk assessment of the AI system in-isolation (TrustGuard) and in-context (SSM) along with a set of recommended countermeasures to mitigate any high impact security risks that are identified. The FAITH and SSM Knowledgebases that underpin these semi-automated risk assessment tools have been jointly developed to include the full range of the latest security threats, including those that are specific to LLM and Agentic AI systems, and to systems with multiple linked AI models. In this way, FAITH supports consistency with the NIST MEASURE 2.7 function requirements for security assessment and countermeasure identification and implementation.

| NIST MEASURE | FAITH Trustworthiness Attribute (TWA) | FAITH Tool or Service |
|---|---|---|
| Measure 2.7: AI system security and resilience are evaluated | Security & Resilience | TrustGuard<br>SSM (+ Knowledgebase)<br>FAITH Knowledgebase |

### 2.2.8 Measure 2.8: Risks associated with transparency and are examined

This involves enabling meaningful visibility into entire AI pipelines, workflows, processes or organisations in order to minimise the disconnect between AI developers, users and wider affected communities. Such visibility can involve the transparent disclosure of data provenance, model design choices, training procedures, and deployment contexts, thereby allowing stakeholders to understand how AI-enabled decisions are produced and governed. It also allows for increased scrutiny and can empower users and impacted groups to articulate domain-relevant concerns, contest outcomes, and contribute to the refinement of requirements and safeguards. Accountable leadership – whether individuals or groups – and transparent roles, responsibilities, and lines of communication foster and incentivise quality assurance and risk management activities within organisations. Conversely, a lack of transparency will complicate trustworthiness activities. Effective transparency leads to improved trust and legitimacy of AI systems within affected stakeholder groups, resulting in greater technology acceptance and the foundations for participatory, socially attuned, and ultimately sustainable AI development and deployment. This maps directly to the FAITH TWA – Accountability & Transparency.

Key **suggested actions** here include tracking activities (e.g. logging, version histories…etc); co-design activities (e.g. creating expert AI developer, end user and UX/UI and HCI experts teams); ensuring diverse stakeholder needs are accounted for; ensuring human oversight; providing accessible policies that detail accountability processes; and monitoring the effectiveness of those policies. Although the FAITH project tools do not directly involve themselves with the organisational governance of AI systems (e.g. policy-led accountability), the ability to permanently document metadata and risk assessments throughout the AI lifecycle with the **AI Model Passport**, the **Data Passport**, the **FAITH AI Model Hub, TrustGuard, SSM** and the **Risk Report Compiler** provides an evidential base that can be used to support measures of transparency and to inform accountability assessments. Furthermore, the **trustSENSE** tool allows an organisation to assess the maturity of different teams (developers and end users) in relation to developing or using AI systems. It provides an indication of the Responsibility & Ethics and Policy Adherence levels of those teams and as such is, in part, also capable of supporting consistency with some aspects of the NIST MEASURE 2.8 function requirements for transparency and accountability.

| NIST MEASURE | FAITH Trustworthiness Attribute (TWA) | FAITH Tool or Service |
|---|---|---|
| Measure 2.8: Risks associated with transparency and are examined | Accountability & Transparency | trustSENSE<br>AI Model Passport<br>Data Passport |

| | | |
|---|---|---|
| | | FAITH AI Model Hub<br>TrustGuard<br>SSM<br>Risk Report Compiler |

### 2.2.9 Measure 2.9: The AI model is explained, validated, and documented, and AI system output is interpreted within its context

This involves supporting users of AI systems in gaining a deeper or clearer understanding of the processes by which an AI system produces its outputs, as well as helping users to interpret those outputs effectively. When users possess a clearer grasp of underlying processes, including data dependencies, model assumptions, and key decision pathways, they are better positioned to assess the appropriateness and limitations of AI-generated results in their specific context. Such understanding strengthens users' ability to detect anomalies, challenge questionable outputs, and integrate domain expertise into final decisions rather than deferring uncritically to algorithmic authority. Equally important is assisting users in interpreting outputs through intelligible explanations that align with their technical and contextual knowledge. This enables more proportionate trust, more nuanced human–AI collaboration, more meaningful human oversight, and for the socially legitimate deployment of AI technologies. This maps directly to the FAITH TWA – Explainability & Interpretability.

Key **suggested actions** here include designing-in explainability systems from the outset; using established and well-understood approaches; co-design and testing of various and different explanations and explanation approaches with effected stakeholders; documenting AI model features and metadata with data statements and model cards; and provide explanations of system characteristics according to defined properties or attributes. FAITH directly operationalises this through the **AI Model Passport**, the **Data Passport** and the **FAITH AI Model Hub** which both manually and automatically captures and makes easily accessible AI model and dataset features and metadata information in the form of model cards. In addition, the **FAITH AI Trustworthiness Assessment Framework** defines the 7 Trustworthiness Attributes (TWAs) that have been referred to in this Chapter, which are designed to enable easy interpretation of the trustworthiness of an AI system against those properties and attributes. In this way, FAITH supports consistency with the NIST MEASURE 2.9 function requirements for explainability of AI model and dataset features, metadata, model cards and easily-interpretable trustworthiness attributes.

| **NIST MEASURE** | **FAITH Trustworthiness Attribute (TWA)** | **FAITH Tool or Service** |
|---|---|---|
| Measure 2.9: The AI model is explained, validated, and documented, and AI system output is interpreted within its context | Explainability & Interpretability | AI Model Passport<br>Data Passport<br>FAITH AI Model Hub<br>FAITH AI TAF |

### 2.2.10 Measure 2.10: Privacy risk of the AI system is examined

This involves ensuring that privacy values, such as anonymity, confidentiality, dignity, informed consent and control inform the development and deployment of AI systems in order to safeguard stakeholders. Privacy values should function as normative constraints on data collection, processing, and utilisation (inc. training), thereby limiting the potential for intrusive surveillance, discrimination, or other forms of harm. Prioritising dignity requires that AI applications avoid exploitative or dehumanising uses of data, especially in contexts involving vulnerable populations. Informed consent and meaningful user control over data and interactions ensure that individuals retain agency regarding

how their information is used and how AI-mediated decisions affect them. The inclusion of Privacy-enhancing Technologies (PETs) for AI systems help establish robust safeguards against misuse, support accountability, and promote equitable distributions of risk and benefit among all affected stakeholders. This maps directly to the FAITH TWA – Privacy Enhancement.

Key **suggested actions** here include defining the important privacy aspects through co-design with end users and other affected stakeholders; maintaining and regularly reviewing compliance with all relevant privacy regulations (e.g. GDPR); applying data anonymisation techniques where appropriate; and ensuring access controls to training and production datasets are in place. FAITH directly operationalises this through the **SSM** tool, which includes a risk assessment of privacy and compliance threats. Within the SSM Knowledgebase are encoded GDPR compliance threats and there is provision within the modelling interface to define datasets with different sensitivities (e.g. biometric data; genetic data; health data; sensitive data; and special category data). Hence compliance threats and their associated mitigation and control strategies can be readily assessed. In this way, FAITH supports consistency with some aspects of the NIST MEASURE 2.10 function requirements for privacy compliance with regulations.

| NIST MEASURE | FAITH Trustworthiness Attribute (TWA) | FAITH Tool or Service |
|---|---|---|
| Measure 2.10: Privacy risk of the AI system is examined | Privacy Enhancement | SSM |

### 2.2.11 Measure 2.11: Fairness and bias are evaluated

This involves paying attention to the concepts of equality and equity, including the identification and mitigation of harmful biases and discriminatory outcomes, as well as the nuances of fairness, as related to the social context in which an AI system is deployed. Equality requires that individuals are treated consistently, while equity recognises that different groups may need differentiated treatment to achieve just outcomes. In AI contexts, this demands the systematic identification and mitigation of harmful biases embedded in data, models, and workflows, as these can lead to discriminatory outcomes for marginalised or historically disadvantaged groups. Beyond technical de-biasing, careful consideration of fairness involves engaging with the specific social, cultural, and legal context in which an AI system operates, including power asymmetries and structural forms of exclusion. Nuanced notions of fairness may vary across domains, stakeholders, and jurisdictions, and thus are challenging to be reduced to a single metric or optimisation target. This maps directly to the FAITH TWA – Fairness.

Key **suggested actions** here include: (i) the assessment of the computational and statistical forms of bias, (ii) the identification of who could be harmed and what those harms could be within specific contexts, (ii) the examination of within-group and intersectional disparities, (iii) the analysis of bias sources in both training and TEVV datasets, (iv) the assurance of representativeness, balance and minimal systemic distortions in multimedia datasets, (v) the engagement of affected stakeholders in impact assessment, and (vi) the establishment of acceptable levels of fairness and policies and actions to ensure these are not exceeded. Initially, FAITH operationalises these through the **TrustGuard** and **SSM** tools. Firstly, TrustGuard enables the definition, execution, and monitoring of fairness metrics and thresholds tailored to the context of use, while SSM integrates fairness-related risks into the broader socio-technical risk assessment. For example, SSM is capable of providing specific risk assessment on data assets, such as the potential for a loss of fairness in the output dataset from an AI model. This means SSM is capable of indicating a lack of equality or equity in the data, with issues such as harmful bias or discrimination being present, resulting in unfair predictions made by the AI model. It will also recommend controls to introduce to mitigate this loss of fairness. In addition, the DBDM

(Data Bias Detection and Mitigation) toolkit[3], a FAITH-developed enabler of MEASURE 2.11, is designed to support the systematic detection and mitigation of data-level bias prior to model training. It is part of the AI model passport and it has been developed by FORTH to: (i) assess bias using a diverse set of information-theory related and statistical-based metrics, and (ii) mitigate bias using synthetic data and graph-based approaches to populate the under-represented groups. In addition, it supports cluster-based subgroup analysis to uncover biases that may affect specific subgroups or intersectional populations. Through this toolkit, bias detection at the data level is systematically linked to model-level fairness assessment and risk management activities. In this way, FAITH supports consistency with some aspects of the NIST MEASURE 2.11 function requirements for fairness and bias evaluations.

| NIST MEASURE | FAITH Trustworthiness Attribute (TWA) | FAITH Tool or Service |
|---|---|---|
| Measure 2.11: Fairness and bias are evaluated | Fairness | TrustGuard<br>SSM<br>AI model passport (DBDM toolkit) |

### 2.2.12 Measure 2.12: Environmental impact and sustainability of AI model training and management activities are assessed

This involves assessing the level of environmental impact resulting from the large-scale, high-performance computational resources used when training and operating AI systems, as well as from the manufacture, maintenance and disposal of the hardware and infrastructure required to operationalise AI systems. Large-scale, high-performance computing used for training and operating advanced models can entail significant energy consumption and associated greenhouse gas emissions, often concentrated in energy-intensive data centres. Evaluating these impacts is essential for aligning AI development with climate and sustainability commitments and for informing decisions about model architecture, training frequency, and deployment strategies. Equally critical are the environmental burdens linked to the manufacture, maintenance, and disposal of specialised hardware and supporting infrastructure, which involve raw material extraction, complex global supply chains, and e-waste management challenges. This does not map to any of the FAITH TWAs and there are no FAITH tools or services capable of assessing the environmental risks associated with an AI system. Assessing environmental impact risks are beyond the scope of the FAITH project because it is a hugely complex area requiring specialist risk assessment tools and approaches, such as environmental impact indicators, energy consumption, efficiency and GHG emissions measures, calculation of 'rebound effects' (where efficiency gains are offset by increased resource consumption), supply chain behaviours, manufacturing processes and so on.

### 2.2.13 Measure 2.13: Effectiveness of the employed TEVV metrics and processes in the MEASURE function are evaluated

This includes critically reviewing and assessing the processes, tools and measures developed to support the validation activities in Measures 2.1 to 2.12 for implicit or systemic bias, unintended consequences, inadvertent oversimplification, potential for gaming, and a lack of nuance. Without such scrutiny, validation frameworks may embed implicit or systemic biases, reproducing inequities under the guise of scientific rigour and compliance. Oversimplified metrics, benchmark datasets, and evaluation protocols can fail to capture contextual nuances, thereby masking disparate impacts on different social groups or application domains. Furthermore, rigid or narrowly defined measures may invite strategic "gaming," where systems are optimised to perform well on specific tests while remaining unsafe, unfair, or unreliable in real-world settings. Systematic meta-evaluation of validation

[3] https://github.com/FAITH-FORTH/DBDM.

practices promotes the incorporation of interdisciplinary perspectives, including ethical, legal, and sociotechnical considerations, into what might otherwise be treated as purely technical exercises. Ultimately, reflective assessment of validation processes is a precondition for trustworthy AI governance, as it strengthens the robustness, legitimacy, and accountability of claims that AI systems have been adequately tested and found fit for purpose.

Key **suggested actions** here include reviewing metrics and TEVV processes for errors; usefulness; usability; appropriacy; acceptability; and effectiveness. FAITH directly operationalises this through the FAITH Tools Validation Methodology that will be described in Chapter 3 of this deliverable.

## 2.3 Summary

Across all aspects of the FAITH project output, the relationship between the FAITH TWAs and the FAITH tools and services to supporting the validation of AI systems according to the NIST AI RMF MEASURE function sub-measures has been clearly demonstrated. Eleven of the thirteen sub-measures are in whole or in part supported by FAITH outputs, as summarised in the Table below.

*Table 1. Mapping between NIST AI RMF MEASURE sub-functions (2.1–2.13) and the FAITH Trustworthiness Attributes (TWAs) together with the FAITH tools/services that support their evaluation. N/A indicates that the FAITH ecosystem does not directly support the corresponding measure, while None denotes that no specific FAITH tool addresses it.*

| NIST MEASURE | FAITH Trustworthiness Attribute (TWA) | FAITH Tool or Service |
|---|---|---|
| 2.1: Documentation | Validity & Reliability<br>Transparency & Accountability | AI Model Passport<br>Data Passport<br>FAITH AI Model Hub<br>TrustGuard<br>SSM<br>Risk Report Compiler |
| 2.2: Human Safety and Representation | Safety<br>Fairness | trustSENSE |
| 2.3: Mixed-method in-context testing | Validation & Reliability | SSM |
| 2.4: The functionality and behavior of the AI system and its components are monitored when in production | Validation & Reliability<br>Safety<br>Accountability & Transparency | N/A |
| 2.5: The AI system to be deployed is demonstrated to be valid and reliable | Validation & Reliability | SSM |
| 2.6: AI system is evaluated regularly for safety risks | Safety | TrustGuard<br>SSM<br>Risk Report Compiler |
| 2.7: AI system security and resilience are evaluated | Security & Resilience | TrustGuard<br>SSM (+ Knowledgebase)<br>FAITH Knowledgebase |
| 2.8: Risks associated with transparency and are examined | Accountability & Transparency | trustSENSE<br>AI Model Passport<br>Data Passport<br>FAITH AI Model Hub<br>TrustGuard<br>SSM<br>Risk Report Compiler |

| | | |
|---|---|---|
| 2.9: The AI model is explained, validated, and documented, and AI system output is interpreted within its context | Explainability & Interpretability | AI Model Passport<br>Data Passport<br>FAITH AI Model Hub<br>FAITH AI TAF |
| 2.10: Privacy risk of the AI system is examined | Privacy Enhancement | SSM |
| 2.11: Fairness and bias are evaluated | Fairness | TrustGuard<br>SSM<br>AI Model Passport (DBDM toolkit) |
| 2.12: Environmental impact and sustainability of AI model training and management activities are assessed | N/A | None |
| 2.13: Effectiveness of the employed TEVV metrics and processes in the MEASURE function are evaluated | N/A | FAITH Tools Validation Methodology |

# 3. FAITH Tools Validation Methodology

## 3.1 Overview of FAITH Tools Validation Approach

As mentioned in Section 1.1, validation takes on a variety of meanings within the FAITH project, thus requiring a multi-tiered approach to validation activities, as will be described in the following sections and summarised in the Figure 2 below:

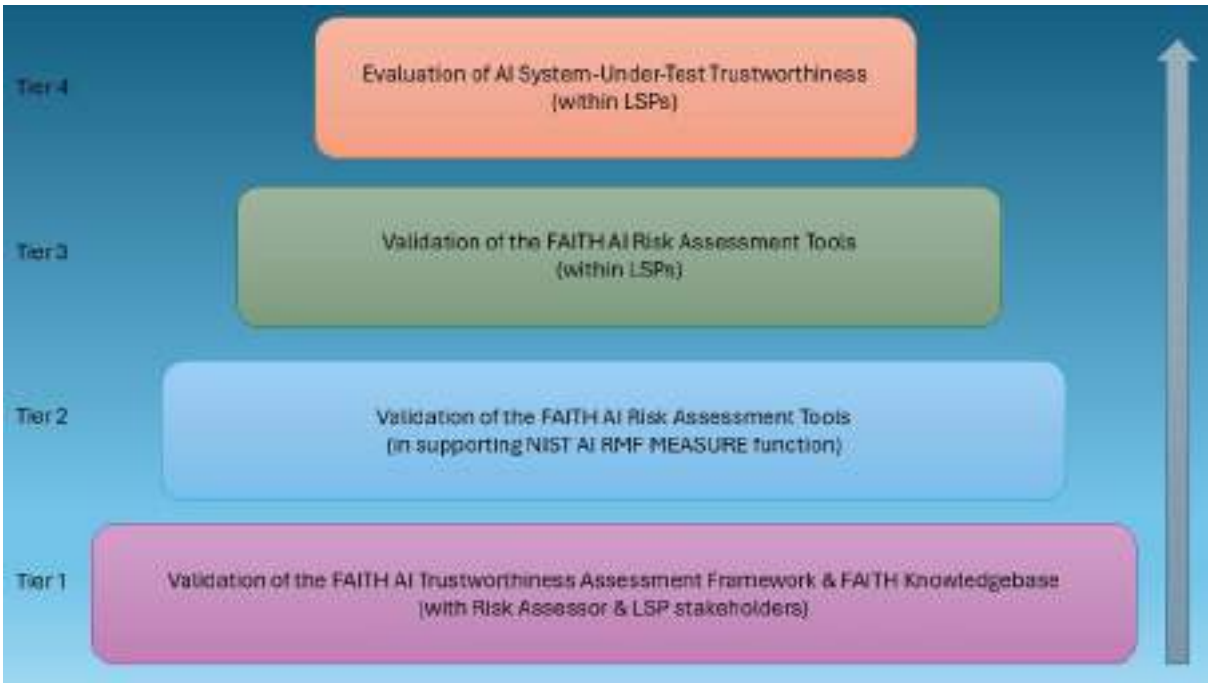



*Figure 2: FAITH Validation Approach Overview.*

## 3.2 Validation of the FAITH AI Trustworthiness Assessment Framework and the FAITH Knowledgebase

The first Tier of the FAITH Validation Approach involves activities to assess the underlying foundational aspects of the FAITH output, namely the FAITH AI Trustworthiness Assessment Framework (TAF) and the FAITH Knowledgebase (KB). Firstly, the FAITH TAF was developed in two phases within in WP2 (v1 and v2), and validation activities have been reported in D2.1 and will be reported in D2.2. In D2.1 the proposed validation methodology was presented, involving the need to assess the main TAF principles, phases and actions with domain and risk experts through a series of validation workshops. A broad outline of the possible validation workshop activities was reported, as well as reports on the initial validation workshop carried out for TAF-v1 (see Section 6.1 in D2.1). This outline plan will be refined, finalised and reported in D2.2, and the workshops themselves will be conducted through M28-M32, with feedback incorporated into final TAF-v2 updates thereafter.

Secondly, the validation and expansion of the FAITH Knowledgebase (KB) constituted a central activity within WP2, ensuring that the core FAITH risk assessment tools are grounded on real-world threats as specifically relevant to the real-world deployment conditions across all LSPs. While the initial structure and content of the FAITH KB were derived from authoritative sources such as ENISA's AI cybersecurity threat landscape[4] , the ISO27000x series, the NIST's AI RMF, and other established sources, it was explicitly conceived as a living repository whose relevance and completeness would be iteratively validated and enhanced through systematic engagement with LSP stakeholders.

[4] https://www.enisa.europa.eu/publications/enisa-threat-landscape-2025

To operationalise this validation, FAITH adopted a structured co-design journey with the LSPs, in order to validate and enrich the KB with domain-specific knowledge on threats, vulnerabilities, consequences, and control measures relevant to AI-based systems. The overarching objective of this co-design activity was to enhance the FAITH KB so that it reflects not only generic AI risks, but also the concrete socio-technical challenges encountered in real deployment environments.

The first stage of KB validation focused on assessing the adequacy of the baseline KB structure in capturing trustworthiness risks across heterogeneous domains. As described in D2.3, the KB is implemented as a relational model interlinking AI lifecycle phases, AI assets, threats, vulnerabilities, controls, consequences, and TWAs. Early dry-run exercises and preparatory discussions with LSP partners revealed that, while the baseline KB provided strong coverage of generic risks, several pilot-specific vulnerabilities and contextual risk factors were insufficiently represented. This finding motivated a deeper, participatory validation process centred on the LSPs themselves, recognising them as primary knowledge holders of domain-specific risks, operational constraints, and regulatory obligations.

The co-design journey followed a clearly defined sequence of activities, comprising: (i) planning of co-design workshops, (ii) execution of LSP-specific workshops, (iii) systematic identification of vulnerabilities, threats, consequences, and control measures, (iv) final assessment of proposed KB extensions, and (v) structured integration of validated inputs into the FAITH KB. This process was carried out over a defined timeline, with dedicated workshops conducted with each LSP between April and May, ensuring full coverage of all pilot domains before final KB consolidation.

Building on this approach, Task 2.3 activities reported in D2.5 operationalised KB validation as a structured co-creation process with LSP stakeholders (see Figure 3). Dedicated KB enhancement templates were provided to all pilots to ensure consistency and methodological rigour across domains. These templates guided participants to identify new vulnerabilities, associated threats, applicable control measures, affected AI assets, lifecycle stages, and relevant trustworthiness attributes. This structured elicitation ensured that all contributions could be seamlessly aligned with the existing KB schema, preserving internal consistency while enabling meaningful expansion.

The co-design workshops played a critical role in validating the practical applicability of the KB. LSP stakeholders were encouraged to reflect on real deployment scenarios, organisational practices, and regulatory contexts specific to their domains, including healthcare, active ageing, media, education, transportation, robotics, and industrial systems. The work focused on understanding how AI systems are embedded within each LSP domain, how they interact with data sources, human actors, and operational processes, and how these interactions shape trustworthiness risks. For each LSP, domain experts and technical partners collaboratively identified key threats by analysing the corresponding AI SUT across the AI lifecycle, including data acquisition, model design and training, deployment, monitoring, and use in decision-making. This process explicitly considered domain-specific constraints, such as regulatory obligations, safety requirements, data modalities, levels of automation, and the role of human oversight. This led to the identification of domain-specific risks not fully captured in the baseline KB, such as context-dependent bias mechanisms, human–AI interaction failures, safety-critical operational constraints, and compliance-driven vulnerabilities linked to sectoral regulation (e.g. GDPR, MDR, AI Act). Importantly, these inputs extended beyond purely technical considerations, reinforcing the KB's socio-technical orientation.

In parallel, the identification of threats was systematically linked to understanding which FAITH Trustworthiness Attributes (TWAs) are most critical within each domain. Rather than assuming equal importance for all TWAs, the process highlighted that attributes such as safety, robustness, privacy, fairness, explainability, accountability, and reliability vary in relevance depending on the use-case context and the potential impact of failure. This alignment between threats and TWAs not only

informed the prioritisation of risks and the selection of appropriate controls, but also served as a further validation activity for the FAITH TAF-v1, where the TWAs were enunciated.

As a result, a range of context-dependent risk patterns, including but not limited to data quality and representativeness issues, robustness and generalisability limitations, explainability and transparency gaps, privacy and compliance risks, and, where applicable, generative AI–specific threats such as prompt-based attacks and hallucinations, were identified and incoporated into the extended FAITH KB. Hence, a key outcome of the validation activity was the enrichment of the KB with pilot-specific threat–vulnerability–control mappings. For each LSP, validated entries were consolidated following a final assessment step and subsequently integrated into the FAITH KB as enhanced or new records. This ensured that the KB could support context-aware risk assessment within TrustGuard, allowing assessments to reflect real-world deployment conditions rather than relying solely on generic AI risk taxonomies.

Beyond content expansion, the co-design journey also functioned as a primary formative evaluation. Feedback gathered during and after the workshops informed refinements in terminology, clarification of definitions, and improved alignment between TWAs and stakeholder perceptions of risk. This was particularly relevant in domains where trustworthiness is primarily understood through human impact, accountability, and transparency rather than conventional cybersecurity metrics.

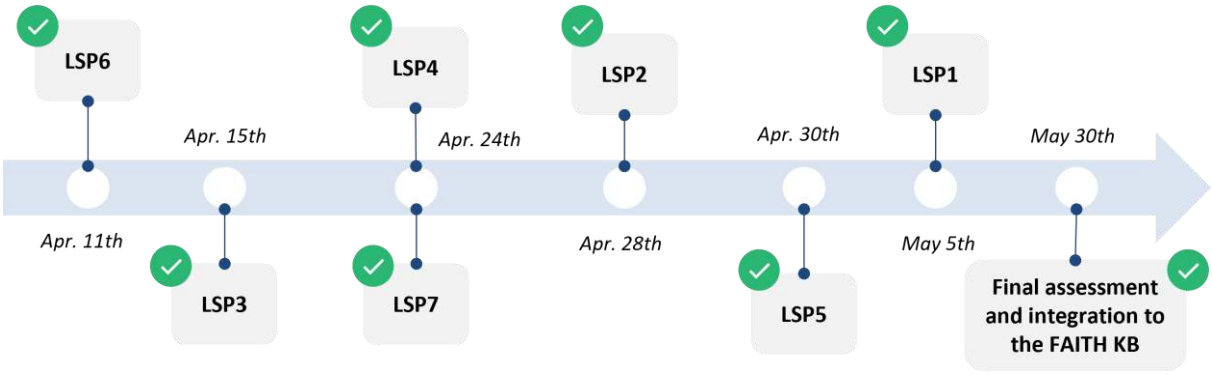



*Figure 3: The KB co-design journey with the LSPs.*

In summary, the FAITH KB validation activity transformed the KB from a standards-based reference repository into a context-aware, pilot-validated knowledge asset. Through a structured, time-bound co-design journey with all LSPs, the KB was systematically expanded, validated, and consolidated, ensuring coverage of domain-specific risks while maintaining alignment with international standards and European regulatory frameworks. This process provides a robust foundation for subsequent FAITH Tool validation activities and supported the long-term evolution of the FAITH TAF-v2.

## 3.3 Validation of the FAITH Tools Against NIST AI RMF MEASURE Function

This Tier of the validation plan involved mapping the FAITH Tools against the NIST AI RMF MEASURE functions in order to demonstrate how FAITH outputs can support the evaluation of AI Systems according to the recommended criteria defined in the NIST source document. This has been extensively reported in Chapter 2 above.

## 3.4 Mixed-methods FAITH Tool Validation Plan

This Tier of the validation approach involves the deployment of the FAITH risk assessment tools into the LSP contexts, with initial deployment focusing on In-isolation risk assessments (using the AI Model & Data Passports, the AI Model Hub, trustSENSE and TrustGuard), followed by the In-context risk assessments (using SSM). A structured process of pre-validation (dry run testing), deployment and use within LSPs, and formal, standardised, structured feedback has been developed and implemented.

### 3.4.1 Dry Run Testing

The Dry Run Testing phase constitutes a structured pre-operational validation stage conducted prior to the full deployment of project solutions across all Large-Scale Pilots (LSPs). Given the cross-sectoral nature of the project, this phase is designed to ensure that TrustGuard, the in-isolation risk assessment system, will meet the needs of all heterogeneous environments accommodated by FAITH through the various LSPs. The dry run, operationalized the FAITH-TAF-v01 methodology in a spreadsheet format, and, then, it was distributed to the pilot partners to test the quality, relevance, and answerability (based on end-user's expertise) of the questions posed to the end-user of TrustGuard. This allowed systematic validation, issue identification, and corrective refinement before the actual development and deployment of TrustGuard.

Within this phase, the dry-run was deployed in pilot-specific staging environments and integrated with the respective LSP infrastructures, including AI systems at all lifecycle stages, from development to deployment. The objective is to validate applicability across diverse sectoral architectures, ensuring that, eventually, TrustGuard will be able to ingest relevant system information, execute trustworthiness risk assessments, and generate consistent outputs regardless of domain-specific variations. This step confirms the portability and adaptability of the tool across healthcare, media, transportation, education, active ageing, wastewater management, submarine drones, or any other participating LSPs. Representative AI systems within each LSP are assessed to verify proper threat identification, vulnerability evaluation and impact estimation. Particular attention is given to confirming that sector-specific contextual parameters are correctly captured. This cross-pilot validation ensures methodological consistency while allowing contextual flexibility.

Identified issues, corrective actions, and pilot-specific observations were documented, leading to a structured readiness assessment. The outcome of this phase ensures the methodological consistency, operational stability, and regulatory alignment of TrustGuard before transitioning to large-scale production environments. The output of this dry run is documented in D2.3.

### 3.4.2 FAITH Tool Validation Plan with LSPs

The first phase of the validation plan involved developing detailed implementation plans with each of the LSPs. These implementation plans have defined when and in what order the LSPs should engage with the FAITH In-isolation risk assessment tools. It starts by defining the training and support provision provided by the tool developers and then defines the timings and key stakeholders for using and feeding back on each of the FAITH tools. It then goes on to define the actions, training and engagement & feedback timings for the second In-context phase of the validation plan.

These plans have been developed with each LSP in group and individual sessions to reflect realistic timeframes that are appropriate for each specific deployment context. They form the basis for activity planning and progress reporting.

### 3.4.3 Formal Feedback and Iteration

A critical part of the implementation planning has been the development of a standardised feedback process capable of capturing feedback and evaluation of the full range of FAITH tools across all 7 LSP contexts. This process involves a mixed-methods approach using a standardised feedback survey for each tool with open and closed questions, and specific LSP 1-to-1 feedback workshops.

The feedback has been carefully structured to reflect established evaluation and measurement approaches. Firstly, it draws on the Technology Acceptance Model[5] (including TAM2[6] and TAM3[7] extensions) to validate the Usefulness, Ease-of-Use and Social/Human attitudes towards the various FAITH Tools. The Technology Acceptance Model (TAM) and its extensions are used to explain and predict whether people will actually use a technology, and to diagnose what needs to change in a tool or service to increase adoption. It provides a theoretical explanation of how users' perceptions of a system translate into intentions and actual use, via perceived usefulness and perceived ease of use in order to predict likelihood of acceptance or rejection of a given system before or after deployment. By measuring perceived usefulness and ease of use via questionnaires it is possible to identify which aspects of a system or its implementation need improvement (design, training, communication, support). The TAM extensions introduce notions of social influence and personal attitudes and how they may contribute to the likelihood that a technology may be adopted. The TAM model family is used to structure survey instruments and evaluation studies in order to decompose "acceptance" into measurable constructs that guide design decisions, change management, and targeted interventions for specific technologies and user groups.

Within FAITH the TAM measures are operationalised in an online feedback survey consisting of three main sections. Section 1 is focused on the TAM aspects of Usefulness, Ease-of-Use and Social/Human attitudes, which are presented as multiple choice, quantitative, closed questions, the specific wording of which have been tailored to be appropriate to each tool, while the overall question remains simultaneously standardised across tools. Section 2 provides a series of prompted open-question text response qualitative questions designed to support specific detailed and unstructured feedback on any aspect of the tool that the LSP respondents consider important to feedback on. Finally, Section 3 provides a set of multiple choice, quantitative, closed questions, again with minor adjustment to specific tools, reflecting on the extent to which the FAITH tools support aspects of the NIST AI RMF MEASURE function elements. An example survey for one tool can be seen in Appendix A.

Once the survey results have been collected and analysed, a series of 1-to-1 deep-dive interviews or (if preferable for LSPs) group workshops will be conducted to explore in greater depth some of the comments received during the survey phase. These will focus in particular on the detailed written qualitative comments received and will aim to better understand how the various FAITH tools can be specifically adapted and iterated to better suit the full range of LSP domains and contexts. This survey and deep-dive interview process will also be repeated for the In-context phase of the implementation plan.

This process has already been conducted for the trustSENSE tool, where a considerable range and number of feedback and comments were received as a result of an initial trustSENSE training workshop for all LSPs and the feedback collected thereafter, and feedback from LSPs using trustSENSE in deployment contexts. This has led to:

1. Questionnaire updates to better reflect user roles
2. Updates to better reflect FAITH branding
3. Question updates to reduce the use of precise timings and quantities
4. Question updates to focus on routine and recurring practices
5. Question updates to standardise terminology, simplfy complex statements, remove redundant phrasing, make statements more actionable, and to improve inclusivity

[5] Davis, F.D., Bagozzi, R.P. and Warshaw, P.R., 1989. Technology acceptance model. *J Manag Sci*, *35*(8), pp.982-1003.
[6] Venkatesh, V. and Davis, F.D., 2000. A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management science*, *46*(2), pp.186-204.
[7] Venkatesh, V. and Bala, H., 2008. Technology acceptance model 3 and a research agenda on interventions. *Decision sciences*, *39*(2), pp.273-315.

Additional detailed qualitative feedback was received concerning the need for questions to be grouped into related items and separated on the page; for results to be aligned with FAITH TWAs; for both certain questions and the suggested mitigations to be more context-relevant; and for more support in facilitating the use of trustSENSE in deployment settings, especially with end users.

The value of this formalised feedback process and the standardised survey instrument across tools, LSPs and implementation phases is threefold. Firstly, it embeds the validation process into the implementation plan as a critical activity, thereby ensuring a robust validation process. Secondly, it provides detailed formal feedback that will inform the iteration and improvement of the FAITH Tools, including if necessary the ability to further tailor the tools to specific LSP requirements, in preparation for the next phase of implementation during the In-context phase. Thirdly, the standardised approach allows for meaningful comparison across FAITH Tools, enabling an overall evaluation of the whole FAITH offering, not just at the individual tool level.

# 4. Policy Recommendations

## 4.1 Turning Feedback Into Insight

Within Task T3.3, policy recommendations constitute a continuous, project-wide activity that runs throughout the FAITH implementation, rather than a one-off or end-stage exercise. From the early phases of the project, empirical feedback from the LSPs and evaluation activities is systematically collected, analysed, and translated into actionable, evidence-based insights over time. This longitudinal approach ensures that the policy recommendations evolve in parallel with the maturation of the FAITH methodology and tools, remaining aligned with international standards, particularly the NIST AI RMF and ENISA guidance, while being firmly grounded in the operational realities of AI deployment across diverse socio-technical domains.

The evaluation methodology established under T3.3 instantiated the MEASURE function of the NIST AI RMF by combining quantitative, qualitative, and mixed-method approaches to assess AI risks, impacts, and mitigation effectiveness across 13 axes. These axes collectively cover technical performance, safety, security, transparency, accountability, privacy, and fairness, as well as human-centric and organisational considerations. Importantly, the evaluation activities were designed to operate under conditions comparable to real deployment settings, ensuring that findings could meaningfully inform governance and policy-level conclusions.

The primary intended receivers of the policy recommendations emerging from Task T3.3 are European-level advisory and governance bodies concerned with trustworthy AI, with particular emphasis on the High-Level Expert Group on Artificial Intelligence (HLEG). Given HLEG's role in shaping ethical and trustworthiness guidance for AI systems in Europe, the policy recommendations developed within FAITH aim to provide empirically grounded feedback on the applicability, limitations, and operational challenges associated with existing trustworthiness principles and assessment frameworks.

A central mechanism for transforming evaluation outcomes into policy insight is the iterative feedback loop between evaluation results, stakeholder engagement, and the FAITH KB. Evidence collected through pilot evaluations, co-design workshops, and structured assessments are systematically analysed to identify recurring risk patterns, gaps in existing controls, and areas where current regulatory or organisational practices are insufficient to support trustworthy AI deployment. These insights will be then abstracted from individual pilot contexts to identify cross-domain trends and systemic challenges.

At the end, policy recommendations are therefore derived through a bottom-up synthesis process, whereby concrete observations, such as limitations in documentation practices, challenges in demonstrating fairness or explainability, or gaps in human oversight mechanisms, will be mapped to higher-level governance needs.

Crucially, the recommendations reflect the socio-technical nature of AI risk management highlighted throughout FAITH. Feedback from LSP stakeholders demonstrated that trustworthiness cannot be achieved through technical safeguards alone, but requires coordinated policy interventions addressing organisational maturity, accountability structures, human oversight, and lifecycle governance. As a result, the policy recommendations emerging from T3.3 address not only technical requirements (e.g. evaluation frequency, documentation standards, monitoring obligations), but also procedural and organisational measures, such as stakeholder engagement practices, training requirements, and clarity of roles and responsibilities.

The resulting policy recommendations are thus positioned as actionable guidance for policymakers, regulators, and standardisation bodies, supporting the operationalisation of trustworthy AI principles

in line with NIST and ENISA frameworks. By explicitly grounding policy insight in validated evaluation evidence and LSP feedback, FAITH demonstrates how large-scale pilot activities can serve as a bridge between AI risk assessment methodologies and practical, implementable policy guidance.

Based on the current analyses and work carried out, some initial considerations and consequent recommendations can be drawn as follows.

- The regulatory scenario is still really fragmented and there are many initiatives that often mention the same issues or trustworthy dimensions. A common terminology and formal ontological frameworks can help shaping a common understanding.
    - *R1: harmonize terminology and languages:*
        - define a shared glossary across agencies and sectors, aligned with existing standards
        - move towards a lightweight reference ontology for AI risk concepts (hazards, harms, assets, controls, assurance).
        - require new initiatives to map their terms to this ontology to avoid duplication and confusion.
- The language of people working in risk assessment and (cyber-)security regulations is not always straightforward for people working in AI and data science.
    - *R2: work to bridge the language gap between risk/security and AI/data science:*
        - create “translation guides” that pair risk terms (threat, control efficacy, residual risk) with AI terms (model, dataset, evaluation metric, deployment context)
        - Provide plain-language policy profiles with worked examples (e.g., how to do a basic hazard analysis for an LLM-based feature)
        - Fund joint training and cross-functional working groups (risk officers, ML engineers, product, legal) to co-develop templates and checklists
- In many domains, there are already guidelines and standards for risk assessment and ensuring that some key trustworthy dimensions are verified (e.g., reliability/dependability or safety).
    - *R3: tailor (verticalize) frameworks to domains and existing regulations:*
        - issue domain profiles that reference sector standards
        - specify domain-relevant harms, metrics, test protocols, and evidence (assurance cases) rather than generic requirements
        - encourage regulators to accept a common AI risk core plus domain add-ons, reducing redundant audits
- Risk assessment is a practice common in industrial and commercial domains, but less taken into account in academia.
    - *R4: build a risk culture in academia and research:*
        - make basic AI risk assessment and mitigation plans part of grant applications, ethics reviews, and data access approvals
        - provide lightweight, open templates and tooling for research settings (risk registers, model cards, data cards, impact assessments)

- incentivize safe-by-design practices (red-teaming, sandboxing, incident reporting) and include AI risk topics in curricula and lab onboarding
- encourage transparent disclosure of known limitations and evaluation results when releasing models, data, or code

As an overall recommendation:

- *R5: promote interoperable evaluations (benchmark suites, shared testbeds) and incident/near-miss reporting*.

# 5. Conclusions and Next Steps

## 5.1 Overall validation conclusions

The validation activities carried out within Task 3.3 demonstrate that the FAITH framework, tools, and supporting knowledge assets provide a coherent and operationally grounded approach to AI risk assessment aligned with the NIST AI RMF MEASURE function. Through the systematic mapping of FAITH Trustworthiness Attributes (TWAs) and tools to the MEASURE sub-functions (Section 2), the project has shown that eleven out of the thirteen sub-measures are either fully or partially supported by the FAITH ecosystem. This alignment confirms that FAITH does not operate as a standalone assessment toolkit, but as a structured implementation layer capable of translating high-level trustworthiness principles into measurable, documentable, and repeatable validation activities.

The multi-tier validation strategy described in Chapter 3 further reinforces this conclusion. Foundational validation of the FAITH AI Trustworthiness Assessment Framework (TAF) and the FAITH Knowledgebase (KB) ensured that the conceptual and structural underpinnings of the project are robust, standards-aligned, and enriched through domain-specific co-design with all Large-Scale Pilots (LSPs). The KB co-design journey in particular transformed the repository from a standards-derived baseline into a context-aware, pilot-validated knowledge asset. This strengthens the methodological legitimacy of subsequent tool-based risk assessments, as they are grounded in validated socio-technical threat–vulnerability–control mappings rather than purely generic taxonomies.

At the tool level, validation against the NIST MEASURE function has demonstrated that FAITH provides strong support for documentation (Measure 2.1), human-centric evaluation (2.2), mixed-method in-context testing (2.3), safety (2.6), security and resilience (2.7), transparency and accountability (2.8), explainability (2.9), privacy (2.10), and fairness and bias assessment (2.11). The complementary roles of TrustGuard (in-isolation risk assessment), SSM (in-context socio-technical modelling), trustSENSE (stakeholder maturity assessment), the AI Model and Data Passports, and the FAITH AI Model Hub create an integrated validation ecosystem. This layered approach enables quantitative risk scoring, qualitative maturity evaluation, structured documentation, and contextual impact modelling to operate in combination rather than in isolation. As a result, FAITH supports mixed-method validation consistent with NIST's emphasis on socio-technical measurement rather than purely technical benchmarking.

The Dry Run Testing phase and structured deployment planning with LSPs provided additional evidence of operational robustness and cross-domain adaptability. TrustGuard's interoperability across heterogeneous pilot infrastructures, validation of risk logic, and secure-by-design properties were confirmed under controlled but realistic conditions. The formal feedback and iteration mechanism, grounded in the Technology Acceptance Model (TAM) and structured survey instruments, further demonstrated that validation within FAITH is not limited to technical correctness, but also includes usability, acceptance, interpretability, and organisational fit. The early iteration of trustSENSE based on stakeholder feedback illustrates the practical value of embedding evaluation into implementation.

At the same time, the validation process has clarified the boundaries of FAITH's current scope. Real-time production monitoring (Measure 2.4) and environmental sustainability assessment (Measure 2.12) are not directly supported by dedicated FAITH tools. These gaps are explicitly acknowledged and do not undermine the overall validation results; rather, they delineate areas for potential future extension or integration with specialised external tools. Importantly, the explicit identification of unsupported measures reflects methodological transparency and strengthens the credibility of the validation conclusions.

Overall, the evidence gathered across mapping, co-design, dry runs, structured deployment, and formal feedback indicates that the FAITH ecosystem is methodologically sound, standards-aligned, context-aware, and practically deployable. It operationalises key elements of the NIST AI RMF MEASURE function in a structured, repeatable, and documentable manner, while embedding socio-technical considerations and stakeholder engagement into validation activities. Consequently, FAITH can be considered validated at three complementary levels: (i) conceptual robustness of its trustworthiness framework and Knowledgebase, (ii) technical and methodological adequacy of its tools, and (iii) operational applicability across heterogeneous real-world domains. These conclusions provide a solid foundation for future iterations, policy translation, and exploitation of FAITH results in both regulatory and industrial contexts.

## 5.2 Recommendations for future evaluations

Building on the overall validation conclusions and the practical experience gained through dry runs, LSP deployments, and structured feedback activities, several recommendations can be formulated to strengthen future evaluation cycles of the FAITH ecosystem and similar AI risk assessment frameworks.

Future evaluations should move progressively from pilot-controlled environments toward longer-term, lifecycle-spanning validation. While the current validation approach has successfully demonstrated methodological robustness and cross-domain applicability, future work should incorporate longitudinal assessments that track risk evolution over time. This would allow for the systematic evaluation of how trustworthiness profiles change following model updates, infrastructure modifications, regulatory shifts, or organisational restructuring. Embedding periodic reassessment cycles into operational governance processes would further align FAITH with the NIST AI RMF emphasis on continuous measurement and adaptive risk management.

Future evaluations should strengthen integration with real-world monitoring and incident reporting mechanisms. Although FAITH provides structured, repeatable risk assessments at defined lifecycle stages, real-time behavioural monitoring (Measure 2.4) is currently outside its direct scope. Future work could explore interoperability with logging systems, model performance dashboards, or external monitoring tools in order to connect periodic risk assessments with live operational signals such as performance drift, anomalous outputs, or security events. Such integration would enhance the evidence base for dynamic risk recalibration and improve responsiveness to emergent risks.

Evaluation activities should deepen the quantification of socio-technical impacts. The mixed-method approach adopted within FAITH, combining quantitative risk scoring with qualitative maturity and contextual assessments, has proven valuable. However, future evaluations could further refine measurable indicators for attributes such as explainability, human oversight effectiveness, organisational readiness, and fairness perception among affected stakeholders. Developing structured indicator libraries and reference benchmarks per Trustworthiness Attribute (TWA) would support greater comparability across deployments and domains.

The co-design and stakeholder engagement methodology should be institutionalised as a recurring evaluation mechanism. The Knowledgebase (KB) validation journey with LSPs demonstrated that domain experts are essential for uncovering context-dependent vulnerabilities and regulatory nuances. Future evaluations should formalise periodic KB review cycles, enabling domain-specific updates to threats, vulnerabilities, and controls as technologies and regulatory frameworks evolve. This would preserve the KB's status as a living, context-aware asset rather than a static repository.

Further attention should be given to evaluation harmonisation across regulatory frameworks. While FAITH is strongly aligned with the NIST AI RMF MEASURE function, future evaluations should explore structured mappings to additional frameworks such as the EU AI Act, sector-specific safety regulations, and international cybersecurity standards. Developing crosswalk matrices and assurance case

templates would reduce duplication of effort for organisations operating under multiple regulatory regimes and strengthen interoperability of trustworthiness evidence.

Usability and acceptance evaluation should remain a central component of future validation. The structured feedback approach based on the Technology Acceptance Model (TAM) provided actionable insights into usefulness, ease-of-use, and social acceptance dimensions. Future iterations should expand this evaluation to include more granular role-based analysis (e.g., risk managers, developers, compliance officers, end users) and to examine how organisational maturity influences effective adoption. This would ensure that methodological robustness is matched by practical usability and organisational fit.

Future evaluations should explore scalability and automation opportunities. As AI systems grow in complexity and number, manual or semi-manual risk modelling may become resource-intensive. Investigating automated ingestion of model metadata, integration with CI/CD pipelines, or semi-automated updating of system models could improve efficiency while preserving methodological rigour. Such developments would enhance the feasibility of applying FAITH in large organisations or multi-model ecosystems.

Finally, future evaluations should explicitly incorporate meta-evaluation mechanisms. In line with NIST Measure 2.13, evaluation processes themselves should be periodically reviewed for bias, oversimplification, potential gaming, and blind spots. Structured review panels, cross-pilot benchmarking sessions, and external expert audits could serve as mechanisms to ensure that FAITH's validation methods remain fit for purpose, transparent, and credible over time.

## 5.3 Link to subsequent deliverables and exploitation activities

The validation activities reported in this deliverable do not constitute an isolated exercise but form a critical bridge between methodological development, operational deployment, policy translation, and long-term exploitation of the FAITH results. The outcomes of Task 3.3 directly inform subsequent technical refinements, governance-oriented outputs, and dissemination and exploitation actions across the remaining project duration.

The evidence gathered through the mapping to the NIST AI RMF MEASURE function (Chapter 2) and the mixed-method validation activities (Chapter 3) provides a consolidated empirical foundation for the finalisation and stabilisation of the FAITH AI Trustworthiness Assessment Framework (TAF-v2) and the extended FAITH Knowledgebase (KB). The insights derived from LSP co-design workshops, dry runs, and structured feedback loops will be formally integrated into the final versions of the framework and KB in the relevant WP2 deliverables. In this way, D3.3 acts as an evidence-generating layer that strengthens the scientific and operational credibility of the core FAITH methodological assets.

The structured validation findings directly support the refinement and consolidation of the FAITH tools (TrustGuard, SSM, trustSENSE, AI Model Passport, Data Passport, Risk Report Compiler). Feedback collected through TAM-based surveys and deep-dive workshops informs tool iteration, usability improvements, contextual tailoring, and interface adjustments. These refinements will be reflected in subsequent technical deliverables reporting on tool maturation, integration within LSP environments, and final system releases. Thus, D3.3 provides both justification and prioritisation criteria for technical updates in later phases of the project.

The validation results constitute a key input into the final policy-oriented outputs of WP3. The structured transformation of evaluation evidence into policy recommendations (Section 4) is grounded in the empirical findings documented in this deliverable. Subsequent deliverables focusing on dissemination, policy engagement, and stakeholder outreach will leverage these validated insights when interacting with regulatory bodies, advisory groups, and standardisation initiatives. In particular,

the demonstrated alignment with NIST AI RMF and ENISA guidance strengthens the positioning of FAITH as a practical operationalisation framework for trustworthy AI governance.

The validation outcomes enhance the exploitation potential of the FAITH ecosystem. By demonstrating cross-domain applicability across heterogeneous LSPs, methodological robustness, and structured documentation capabilities, the project establishes the maturity level necessary for uptake beyond the pilot environments. The validated toolchain can be positioned for:

- Adoption by public authorities and regulatory sandboxes seeking structured AI risk assessment workflows.
- Integration within organisational AI governance processes in industry sectors such as healthcare, transportation, robotics, and digital services.
- Alignment with compliance and assurance processes under the EU AI Act and related regulatory frameworks.
- Commercial or service-oriented exploitation models, including risk assessment-as-a-service offerings or consultancy-based deployment support.

The validation evidence strengthens dissemination and impact activities by providing concrete, pilot-backed use cases and measurable outcomes. Rather than presenting FAITH as a purely conceptual framework, subsequent communication and exploitation materials can reference validated deployment scenarios, iterative tool improvements, and structured user acceptance results. This enhances credibility when engaging with external stakeholders, research communities, policymakers, and potential adopters.

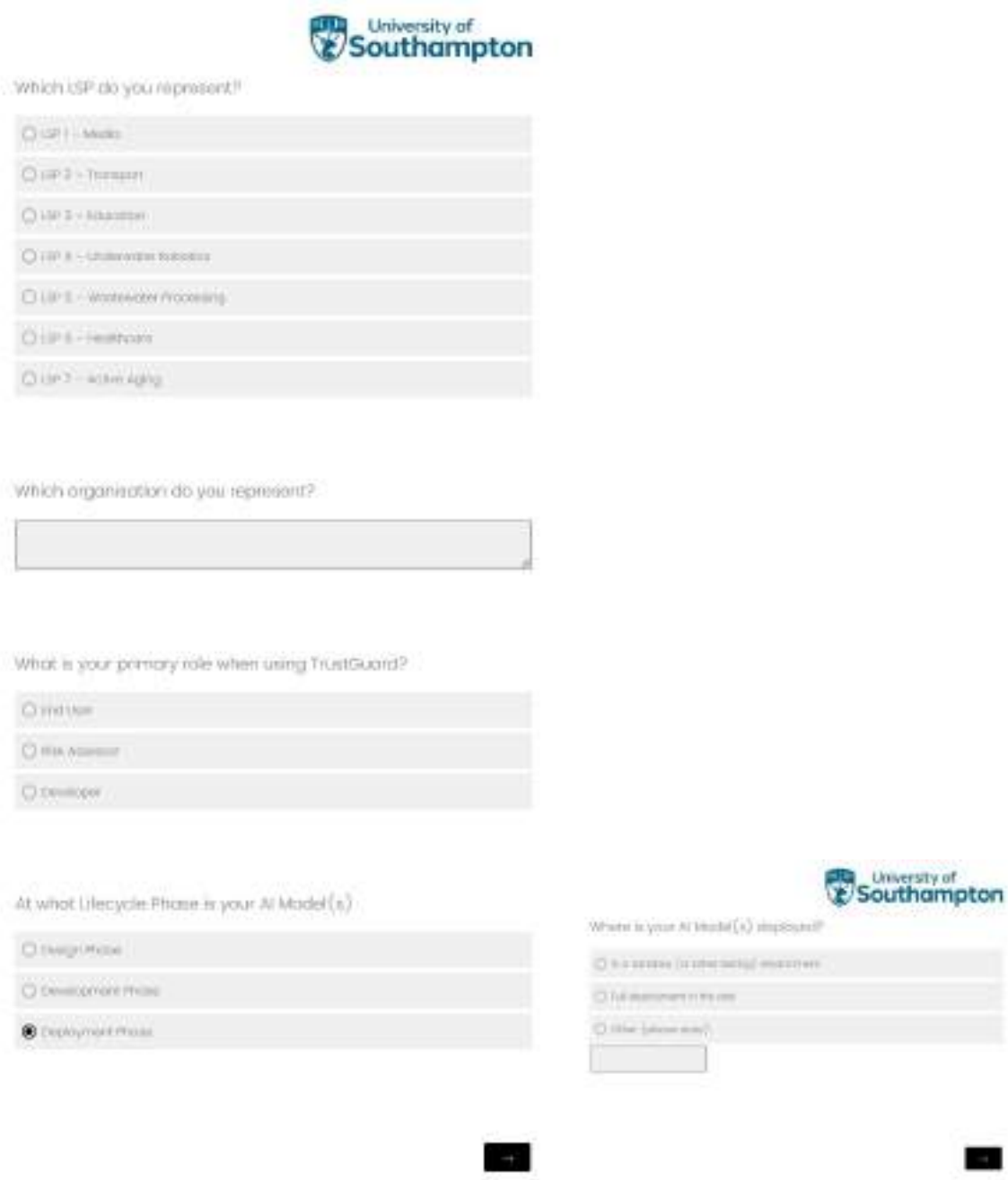
Finally, the structured validation methodology itself represents an exploitable outcome. The mixed-method evaluation plan, including dry-run testing, phased deployment, TAM-based feedback collection, and MEASURE-aligned mapping, can serve as a reusable validation blueprint for future AI governance initiatives. This methodological asset can inform follow-up research proposals, standardisation contributions, or cross-project collaborations within Horizon Europe and beyond.

# Appendix A

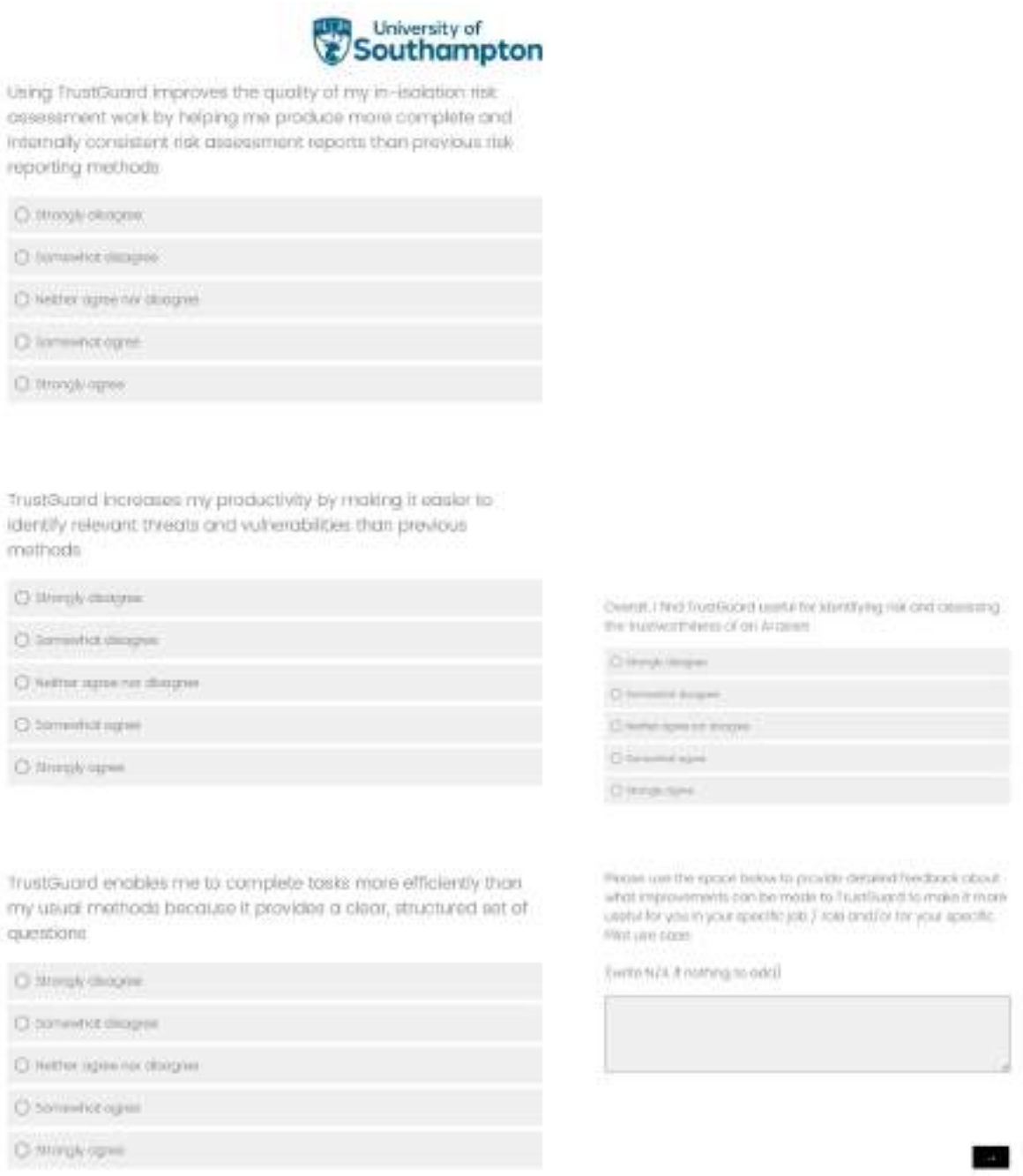
An example of the FAITH standardised Formal Feedback Form.

Please note that this has been developed specifically for the TrustGuard tool and therefore the section concerning the NIST AI RMF MEASUREs is relevant for those sub-measures that are supported by the TrustGuard tool.

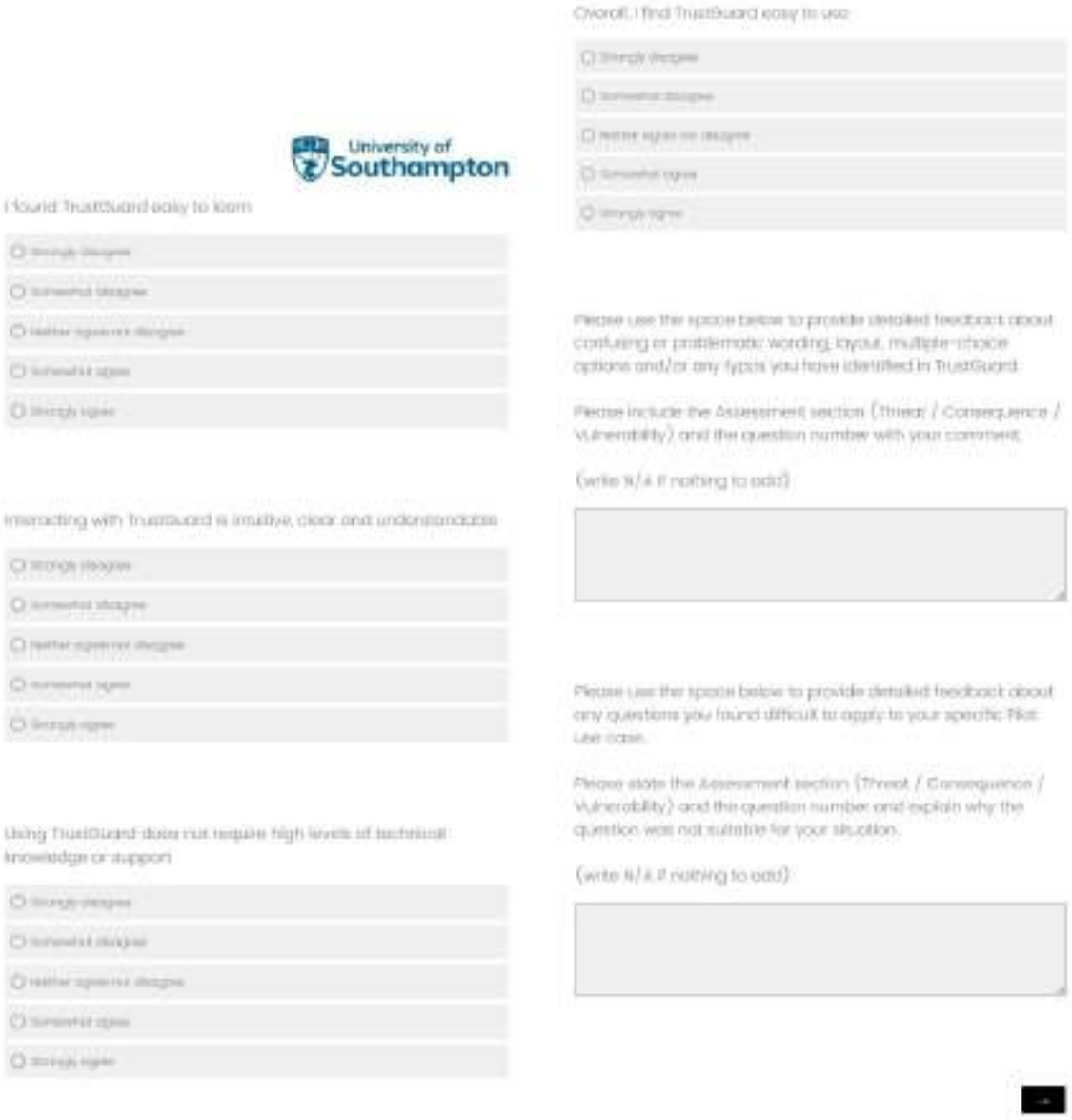
## Section 1: About You



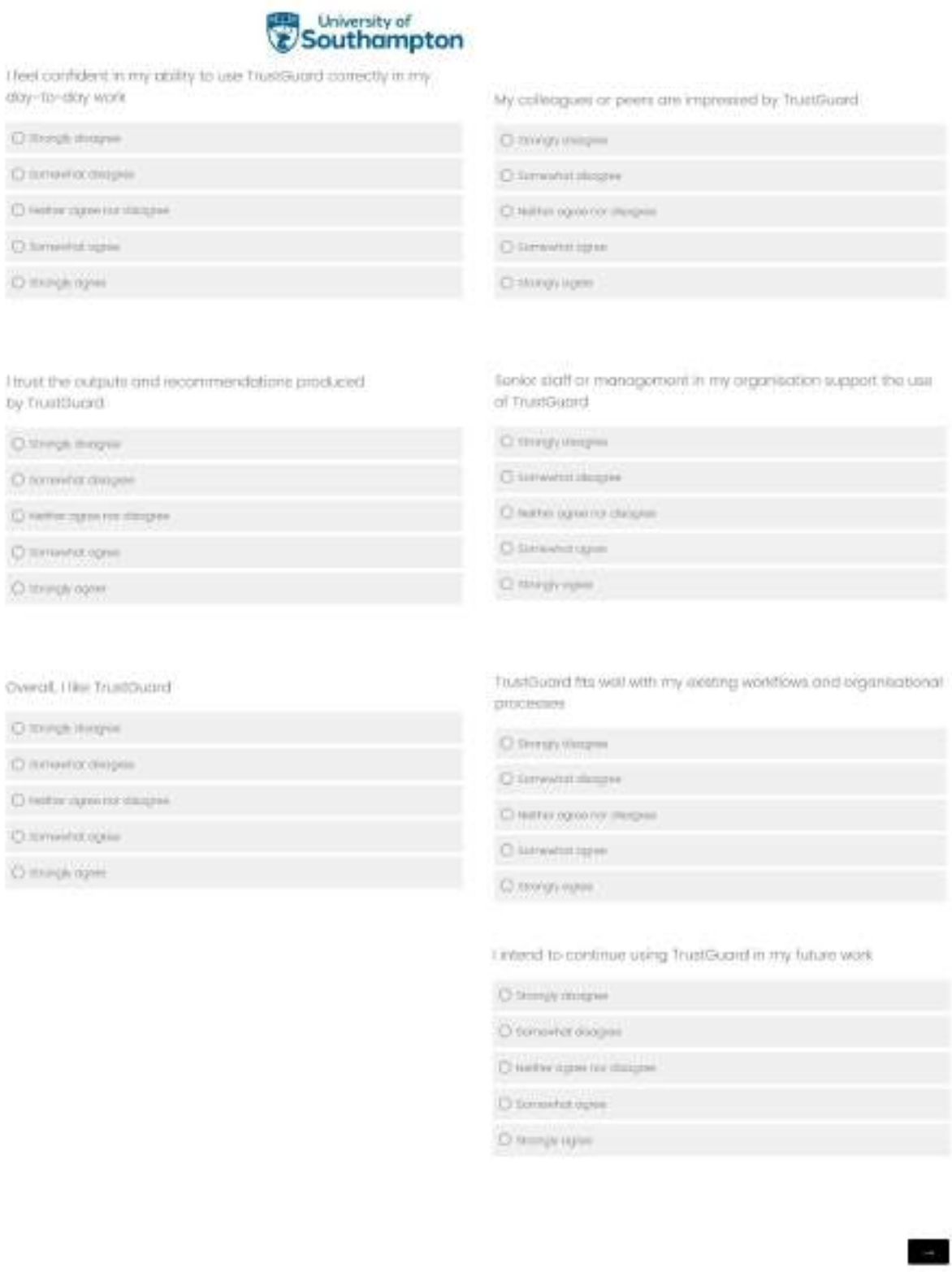
## Section 2: Perceived Usefulness



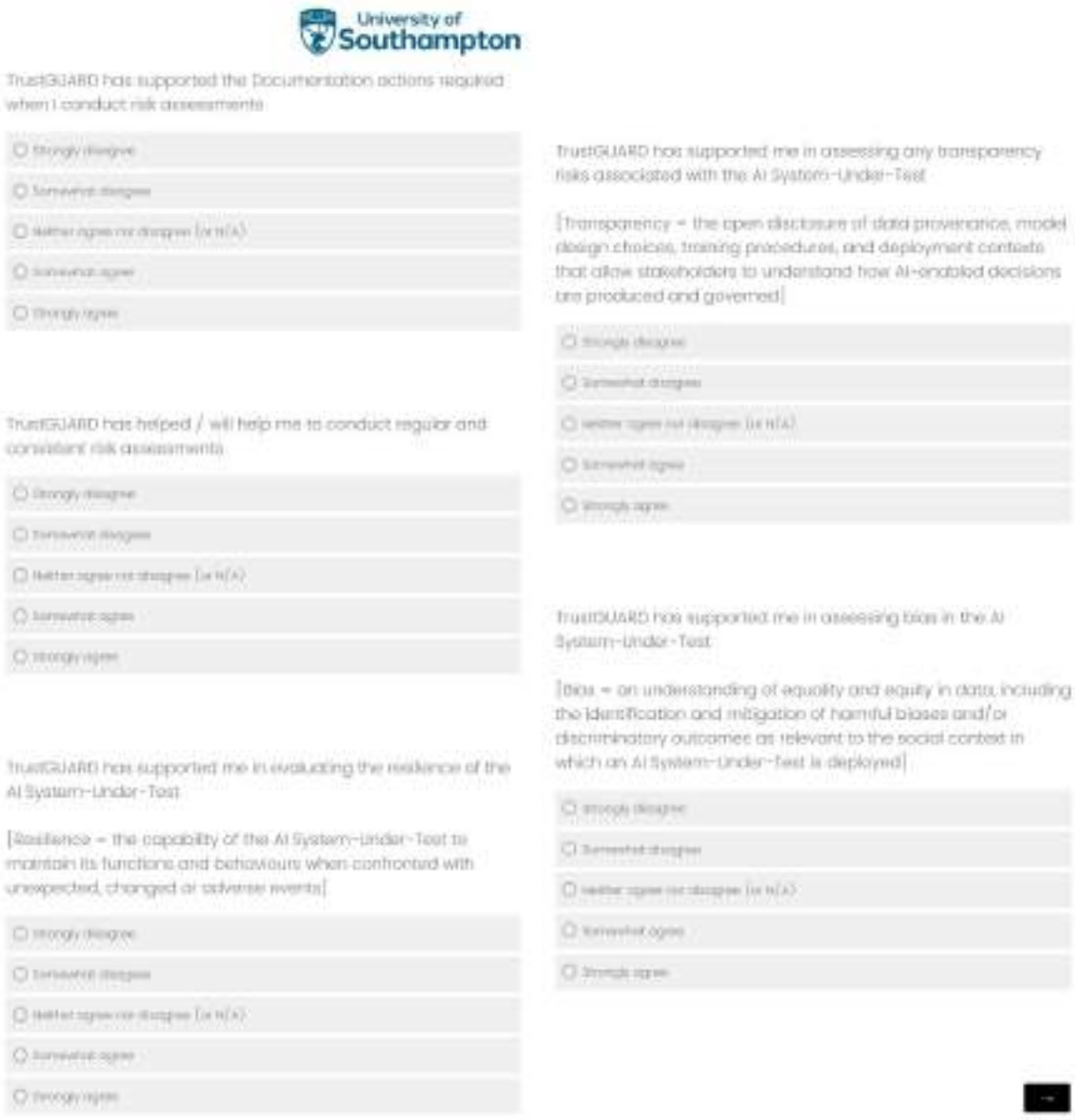
## Section 3: Ease Of Use



## Section 4: Human Factors & Social / Organisation Influences



## Section 5: Support for NIST MEASURES



## Section 6: Detailed Suggestions

