Ref. Ares(2025)5208312 - 30/06/2025

# D2.5 (D13)

## FAITH pilots requirements definition, mapping regulatory regimes and co creation planning

| | |
|---|---|
| **Related Work Package** | **WP2 – FAITH AI Framework for trustworthiness & coordination among Large Scale Pilots (LSPs)** |
| **Related Task** | Task 2.3. - Pilots Requirements Analysis - User Ecosystem Engagement SSH - Scales conditions, measurements-Assessment, KPIs |
| **Lead Beneficiary** | FORTH |
| **Contributing Beneficiaries** | All partners involved in Tasks: 2.3 |
| **Document version** | v.1.0 (by 30/06/2025) |
| **Deliverable Type** | R-Document, Report |
| **Distribution level** | PU-Public |
| **Contractual Date of Delivery** | 30/06/2025 |
| **Actual Date of Delivery** | 30/06/2025 |

| | |
|---|---|
| **Contributors** | Dimitrios I. Fotiadis (FORTH), Kostantinos Marias (FORTH), Michail Klontzas (PAGNH), Nikolaos S. Tachos (FORTH), Vasileios C. Pezoulas (FORTH), Grigoris Kalliatakis (FORTH), Gregoris Mentzas (ICCS), Dimitris Apostolou (ICCS), Evangelia Anagnostopoulou (ICCS), Nikolaos Antonios Grammatikos (ICCS), Sofoklis Sotiriou (EA), Dimitris Koulentianos (EA), Mark Tanner (VPF), Alberto Topini (UNIFI), Asbjørn Følstad (SINTEF), Hilde Johansen (Veas), Sølve Eidnes (SINTEF), Pilar Sala (ACTIVAGE), Héctor Chuliá (ACTIVAGE), Alba Gallego (BRIDG), Gloria Cea (BRIDG), Alfredo Sanz (BRIDG), Maria Antonietta (CNR), Giuseppe R. Leone (CNR), Luca Volpicella (MERMEC), Eleni Tsalapati (ATC), Giorgos Giotis (ATC), Ilias Kanellos (ATC), Thanasis Tsafonis (ATC), Ana Maria Corrêa (KUL). |
| **Reviewers** | Sara Colantonio (CNR), Stefano Modafferi (UoS) |

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 | 14/03/2025 | FORTH |
| **v0.2** | Introduction (Objectives, Relation to the DoA, deliverable structure description) | 20/03/2025 | FORTH |
| **v0.3** | Stakeholder engagement and survey analysis (stakeholder identification and engagement, implementation, survey implementation, sample size estimation) | 10/04/2025 | FORTH |
| **v0.4** | Stakeholder engagement and survey analysis (survey analysis results; overall and subgroups, stakeholder understanding and acceptance of AI) | 22/04/2025 | FORTH |
| **v0.5** | Trustworthy and human-acceptable AI requirements (Key considerations) | 12/05/2025 | ALL LSP leaders |
| **v0.6** | Trustworthy and human-acceptable AI requirements (Applicable legal regimes in LSPs 1-6) | 22/05/2025 | ALL LSP leaders |
| **v0.7** | Co-design workshops (co-design journey, phases of the KB enhancement, outcomes), Conclusions and Future work | 04/06/2025 | FORTH |
| **v0.8** | Review version prepared and shared internally for feedback | 06/06/2025 | ALL LSP leaders |
| **v0.9** | Final comments integrated from reviewers | 26/06/2025 | FORTH |
| **v1.0** | Final version submitted to the European Commission | 30/06/2025 | 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

Deliverable D2.5 focuses on the definition of pilot-specific requirements for trustworthy and human-acceptable Artificial Intelligence (AI) within the FAITH project, while mapping applicable regulatory regimes and outlining the approach to co-creation with stakeholders. It synthesizes the outcomes of Task 2.3, which is dedicated to understanding the real-world deployment context of AI systems in the seven FAITH Large-Scale Pilots (LSPs), through the lens of legal, ethical, operational, and human factors. This deliverable directly supports the iterative refinement of the FAITH Trustworthiness Assessment Framework (FAITH AI_TAF) by aligning stakeholder expectations, domain-specific legal obligations, and trustworthiness criteria. The core part of the deliverable focuses on the methodology and outcomes of the stakeholder engagement activities. These activities included surveys and co-design workshops conducted across the LSPs to identify key concerns, needs, and expectations related to AI systems from users and affected communities. The results provide insights into how stakeholders perceive, understand, and prefer to interact with AI technologies in their respective domains. The analysis captures emerging trends and challenges among stakeholder groups and translates them into practical recommendations to guide developers and implementers. A detailed mapping of the legal and regulatory frameworks that apply to each LSP is provided. These include overarching European legislation, such as the EU AI Act, the General Data Protection Regulation (GDPR), and the Digital Services Act (DSA), as well as national and sector-specific laws relevant to domains like healthcare, education, transportation, media, and smart infrastructure. The deliverable concludes with a synthesis of the trustworthiness requirements and metrics derived from stakeholder inputs and regulatory analysis. These are structured to support the ongoing development and operationalization of the FAITH AI_TAF. The document also outlines the FAITH Knowledge Base (FAITH KB), which serves as a repository for trustworthiness attributes, lifecycle stages, and threat models, enabling each pilot to embed trustworthiness into system design and evaluation. The results of the co-design workshops, including domain-specific enhancements to the FAITH Knowledge Base (e.g. unique vulnerabilities, control measures), are presented to show how co-creation has shaped the project's technical and methodological direction. The FAITH KB is designed as a living, evolving resource which will be continuously refined through iterative stakeholder input, pilot feedback, and legal developments, to ensure that it remains context-aware, practically useful, and aligned with the real-world dynamics of trustworthy AI.

# Table of Contents

1. Introduction ..... 10
1.1. Scope and objectives ..... 10
1.2. Relation to the DoA ..... 10
1.3. Deliverable structure ..... 11
2. Trustworthy and human-acceptable AI requirements ..... 13
2.1. Key considerations and criteria for AI trustworthiness ..... 13
2.2. Applicable Legal Regimes per LSP ..... 14
2.2.1. Applicable Legal Regimes in LSP1 (Related Deliverable: D4.1) ..... 15
2.2.2. Applicable Legal Regimes in LSP2 (Related Deliverable: D5.1) ..... 17
2.2.3. Applicable Legal Regimes in LSP3 (Related Deliverable: D6.1) ..... 18
2.2.4. Applicable Legal Regimes in LSP4 (Related Deliverable: D7.1) ..... 20
2.2.5. Applicable Legal Regimes in LSP5 (Related Deliverable: D8.1) ..... 22
2.2.6. Applicable Legal Regimes in LSP6 (Related Deliverable: D9.1) ..... 24
2.2.7. Applicable Legal Regimes in LSP7 (Related Deliverable: D10.1) ..... 30
3. Stakeholder engagement and survey analysis ..... 33
3.1. Stakeholder identification and engagement strategy ..... 33
3.2. Objectives of the survey ..... 34
3.3. Survey implementation ..... 34
3.4. Sample size estimation ..... 36
3.5. Survey analysis workflow ..... 36
3.6. Survey analysis results ..... 39
3.7. Respondents' understanding of AI ..... 54
3.8. Respondents' acceptance of AI ..... 55
3.9. Respondents' support preferences and interaction with AI ..... 55
4. Co-design workshops to synthesize considerations into AI trustworthiness requirements and metrics ..... 58
4.1. Connection with T2.3 ..... 58
4.2. The co-design journey ..... 59
4.3. The phases of the FAITH KB enhancement process ..... 59
4.4. Supporting material ..... 60
4.5. Outcomes ..... 62
5. Conclusions and Future Steps ..... 78
6. References ..... 80
Appendix ..... 81

A. The FAITH 1st Anonymous Survey ..... 81
B. Detailed survey analysis results ..... 92
B1. Overall ..... 92
B2. Subgroup 1 ..... 98
B3. Subgroup 2 ..... 99
C. The completed KB enhancement templates per LSP ..... 101
C1. LSP1 ..... 101
C2. LSP2 ..... 101
C3. LSP3 ..... 102
C4. LSP4 ..... 102
C5. LSP5 ..... 102
C6. LSP6 ..... 103
C7. LSP7 ..... 103

# List of Figures

**Figure 1.** AI Risk Classification. 19
**Figure 2.** Key steps to an effective quality management system (QMS). 26
**Figure 3.** Steps for CE marking and conformity assessment. 27
**Figure 4.** An instance from the 1st FAITH anonymous survey's confirmation protocol provided by the Research Ethics Committee at FORTH. 33
**Figure 5.** An instance from the EUSurvey ("Editor" panel, Part 1 - Introduction). 35
**Figure 6.** An instance from the EUSurvey ("Editor" panel, Part 2 - Questions). 35
**Figure 7.** An instance from the EUSurvey ("Test" panel). 36
**Figure 8.** The survey analysis workflow. 37
**Figure 9.** Distribution plot of the responses from Q1.1. 39
**Figure 10.** Distribution plot of the responses from Q1.2. 40
**Figure 11.** Distribution plot of the responses from Q1.3. 40
**Figure 12.** Distribution plot of the responses from Q1.9. 41
**Figure 13.** Distribution plot of the responses from Q1.10. 42
**Figure 14.** Distribution plot of the responses from Q1.11. 42
**Figure 15.** Distribution plot of the responses from Q1.16. 43
**Figure 16.** Distribution plot of the responses from Q1.17. 44
**Figure 17.** Distribution plot of the responses from Q1.18. 44
**Figure 18.** Distribution plot of the responses from Q1.19. 45
**Figure 19.** Distribution plot of the responses from Q1.21. 45
**Figure 20.** Distribution plot of the responses from Q1.22. 46
**Figure 21.** Distribution plot of the responses from Q1.23. 46
**Figure 22.** Distribution plot of the responses from Q1.25. 47
**Figure 23.** Distribution plot of the responses from Q1.32. 49
**Figure 24.** Distribution plot of the responses from Q2.1. 49
**Figure 25.** Distribution plot of the responses from Q2.2. 50
**Figure 26.** Distribution plot of the responses from Q2.6. 51
**Figure 27.** Distribution plot of the responses from Q2.7. 51
**Figure 28.** Distribution plot of the responses from Q3.3. 52
**Figure 29.** The structure of the estimated Bayesian model. 53
**Figure 30.** The co-design journey. 59
**Figure 31.** The phases of the FAITH KB enhancement process. 60
**Figure 32.** Useful definitions provided in the FAITH KB enhancement template (Sheet 1). 60
**Figure 33.** Useful consequences provided as an example based on the healthcare LSP (LSP6) in the FAITH KB enhancement template (Sheet 2). 61
**Figure 34.** The core form for the KB enhancement (Sheet 3). 61
**Figure 35.** Drop-down options for the Affected AI lifecycle stage, Affected AI Asset, and Affected TWA dimension/characteristic. 62

# List of Supplementary Figures

**Supplementary Figure 1.** Distribution plot of the responses from Q1.5 (Overall analysis). ..... 92
**Supplementary Figure 2.** Distribution plot of the responses from Q1.6 (Overall analysis). ..... 92
**Supplementary Figure 3.** Distribution plot of the responses from Q1.7 (Overall analysis). ..... 92
**Supplementary Figure 4.** Distribution plot of the responses from Q1.8 (Overall analysis). ..... 93
**Supplementary Figure 5.** Distribution plot of the responses from Q1.12 (Overall analysis). ..... 93
**Supplementary Figure 6.** Distribution plot of the responses from Q1.13 (Overall analysis). ..... 93
**Supplementary Figure 7.** Distribution plot of the responses from Q1.14 (Overall analysis). ..... 94
**Supplementary Figure 8.** Distribution plot of the responses from Q1.15 (Overall analysis). ..... 94
**Supplementary Figure 9.** Distribution plot of the responses from Q1.20 (Overall analysis). ..... 94
**Supplementary Figure 10.** Distribution plot of the responses from Q1.24 (Overall analysis). ..... 95
**Supplementary Figure 11.** Distribution plot of the responses from Q1.27 (Overall analysis). ..... 95
**Supplementary Figure 12.** Distribution plot of the responses from Q1.28 (Overall analysis). ..... 95
**Supplementary Figure 13.** Distribution plot of the responses from Q1.29 (Overall analysis). ..... 96
**Supplementary Figure 14.** Distribution plot of the responses from Q1.30 (Overall analysis). ..... 96
**Supplementary Figure 15.** Distribution plot of the responses from Q1.31 (Overall analysis). ..... 96
**Supplementary Figure 16.** Distribution plot of the responses from Q2.4 (Overall analysis). ..... 97
**Supplementary Figure 17.** Distribution plot of the responses from Q3.1 (Overall analysis). ..... 97
**Supplementary Figure 18.** Distribution plot of the responses from Q3.2 (Overall analysis). ..... 97
**Supplementary Figure 19.** Distribution plot of the responses from Q1.32 (Subgroup 1 analysis). ..... 98
**Supplementary Figure 20.** Distribution plot of the responses from Q2.1 (Subgroup 1 analysis). ..... 98
**Supplementary Figure 21.** Distribution plot of the responses from Q2.2 (Subgroup 1 analysis). ..... 98
**Supplementary Figure 22.** Distribution plot of the responses from Q2.6 (Subgroup 1 analysis). ..... 99
**Supplementary Figure 23.** Distribution plot of the responses from Q2.7 (Subgroup 1 analysis). ..... 99
**Supplementary Figure 24.** Distribution plot of the responses from Q1.32 (Subgroup 2 analysis). ..... 99
**Supplementary Figure 25.** Distribution plot of the responses from Q2.1 (Subgroup 2 analysis). ..... 100
**Supplementary Figure 26.** Distribution plot of the responses from Q2.2 (Subgroup 2 analysis). ..... 100
**Supplementary Figure 27.** Distribution plot of the responses from Q2.6 (Subgroup 2 analysis). ..... 100
**Supplementary Figure 28.** Distribution plot of the responses from Q2.7 (Subgroup 2 analysis). ..... 101
**Supplementary Figure 29.** The completed KB enhancement template of LSP1. ..... 101
**Supplementary Figure 30.** The completed KB enhancement template of LSP2. ..... 101
**Supplementary Figure 31.** The completed KB enhancement template of LSP3. ..... 102
**Supplementary Figure 32.** The completed KB enhancement template of LSP4. ..... 102
**Supplementary Figure 33.** The completed KB enhancement template of LSP5. ..... 102
**Supplementary Figure 34.** The completed KB enhancement template of LSP6. ..... 103
**Supplementary Figure 35.** The completed KB enhancement template of LSP7. ..... 103

# List of Tables

**Table 1.** How D2.5 addresses statements from the DoA. 10
**Table 2.** A summary of the FAITH LSPs. 15
**Table 3.** Relevant legal provisions and technical standards for LSP2. 17
**Table 4.** Mapping of current EU and National legal frameworks relevant to LSP4 operations. 20
**Table 5.** Forthcoming regulatory developments and strategic preparedness for LSP4. 21
**Table 6.** EU MDR software classification requirements relevant to LSP6. 29
**Table 7.** Overview of core GDPR requirements relevant to LSP6. 29
**Table 8.** Key cybersecurity and risk management requirements relevant to LSP6 under EU MDR and MDCG. 30
**Table 9.** Dimensions of stakeholder preferences in human-AI interaction. 57
**Table 10.** Activities connected with T2.3. 58
**Table 11.** The identified unique vulnerabilities, related control measures, corresponding threats, affected AI assets, and trustworthiness (TWA) dimension/characteristics for LSP1. 62
**Table 12.** The identified unique vulnerabilities, related control measures, corresponding threats, affected AI assets, and trustworthiness (TWA) dimension/characteristics for LSP2. 64
**Table 13.** The identified unique vulnerabilities, related control measures, corresponding threats, affected AI assets, and trustworthiness (TWA) dimension/characteristics for LSP3. 66
**Table 14.** The identified unique vulnerabilities, related control measures, corresponding threats, affected AI assets, and trustworthiness (TWA) dimension/characteristics for LSP4. 67
**Table 15.** The identified unique vulnerabilities, related control measures, corresponding threats, affected AI assets, and trustworthiness (TWA) dimension/characteristics for LSP5. 70
**Table 16.** The identified unique vulnerabilities, related control measures, corresponding threats, affected AI assets, and trustworthiness (TWA) dimension/characteristics for LSP6. 72
**Table 17.** The identified unique vulnerabilities, related control measures, corresponding threats, affected AI assets, and trustworthiness (TWA) dimension/characteristics for LSP7. 75

# List of Abbreviations

| Abbreviation | Explanation |
|---|---|
| **AI** | Artificial Intelligence |
| **AI Act** | Artificial Intelligence Act |
| **CEN/CENELEC** | European Standardisation of Standards |
| **CSA** | Cybersecurity Act |
| **CRA** | Cyber Resilience Act |
| **DPIA** | Data Protection Impact Assessment |
| **DPO** | Data Protection Officer |
| **DSA** | Digital Services Act |
| **EDPB** | European Data Protection Board |
| **ENISA** | European Union Agency for Cybersecurity |
| **ETSI** | European Telecommunications Standards Institute |
| **EU** | European Union |
| **FAITH AI_TAF** | FAITH AI Trustworthiness Assessment Framework |
| **GDPR** | General Data Protection Regulation |
| **HLEG** | High-Level Expert Group (on AI) |
| **ISO** | International Organization for Standardization |
| **KB** | Knowledge Base |
| **LSP** | Large-Scale Pilot |
| **MDR** | Medical Device Regulation |
| **NIS/NIS 2** | EU Network and Information Security Directives |
| **NIST** | National Institute for Standards and Technology |

# 1. Introduction

## 1.1. Scope and objectives

Deliverable D2.5 focuses on the identification and analysis of the requirements, perceptions, and challenges related to Trustworthy and Human-Acceptable AI across the seven Large-Scale Pilots (LSPs) of the FAITH project. The deliverable operates within the context of Task 2.3 and aligns with the broader goal of defining operational, legal, ethical, and human-centric conditions for AI deployment and interaction in real-world domains, such as healthcare, education, media. The main objectives of D2.5 include the following: (i) to capture stakeholders' perspectives on AI trustworthiness through multi-method engagement strategies including surveys, co-design workshops, and ethnographic studies, (ii) to assess stakeholder understanding and acceptance of AI, as well as their preferences for AI interaction and collaboration, focusing on domain-specific factors and socio-technical dynamics, (iii) to identify and analyze applicable legal and regulatory regimes for each LSP, integrating sectoral norms (e.g., GDPR, AI Act, MDR, DSA, national legislation) to understand their implications for the AI systems under test (SUT), (iv) to synthesize findings into a set of requirements for Trustworthy AI, (v) to map conflicting stakeholder expectations, (vi) to support co-creation of the FAITH AI Trustworthiness Assessment Framework (FAITH AI_TAF) by aligning technical development with social and legal expectations.

The activities of Task 2.3, which are reported on the current Deliverable D2.5, contribute to several key areas of the FAITH project. The deliverable supports the evolution of the FAITH AI_TAF by incorporating insights from co-design workshops and surveys, enabling refinement of the framework in alignment with the legal, ethical, and technical expectations of LSP stakeholders. In addition, the activities of Task 2.3 focus on the co-design and co-development process of the FAITH Knowledge Base (KB) by integrating empirical data on trustworthiness perceptions, domain-specific risk priorities, and end-user concerns, thereby enhancing the KB's function as a shared, dynamic repository. The analysis of the first anonymous survey provided the baseline of the stakeholders perception on trustworthy AI. Moreover, the findings inform regulatory and ethical integration by aligning user feedback with the legal-ethical risk factors addressed in WP1.

## 1.2. Relation to the DoA

The following table (Table 1) provides the evidence on how the activities of Task 2.3 address the statements of the DoA.

**Table 1.** How D2.5 addresses statements from the DoA.

| Statement from the DoA | How it is addressed in the current deliverable |
|---|---|
| **Task 2.3:** *"This task will identify key considerations & requirements for trustworthy & human acceptable AI over its whole lifecycle."* | **Section 2.1** outlines the key considerations and criteria for AI trustworthiness, incorporating insights from previous deliverables and aligning with European and international standards. |
| **Task 2.3:** *"It will aim to answer some key questions: 'do stakeholders understand the AI?', 'do stakeholders accept the AI?', 'what are the applicable legal regimes for each pilot' and 'how do stakeholders prefer to get support from/partner with the AI?'"* | **Section 3.7** and **Section 3.8** present findings on stakeholder understanding and acceptance of AI systems. **Section 2.2** provides an in-depth mapping of applicable legal regimes per LSP, covering both EU-wide and national-level obligations relevant to each pilot domain. |

| | |
|---|---|
| **Task 2.3:** *"It will work with the pilot stakeholders in the first instance to identify what they perceive to be their main challenges ..."* | **Section 3** presents the results of stakeholder engagement activities, especially through the FAITH 1st Anonymous survey and participatory methods—used to identify perceived challenges, expectations, and trust-related tensions. **Section 4** further incorporates these insights into refined trustworthiness requirements, with attention to conflicting needs across stakeholder groups. |
| **Task 2.3:** *"...to engage with potential stakeholders not directly involved in the project but who could be affected by the AI to assess their perceptions of Trustworthy AI..."* | |
| **Task 2.3:** *"Engagement will be via a combination of techniques such as anonymous crowdsourced survey, focus group discussion, co-design and ethnographic study."* | **Section 3** details the methodology and results from the first round of stakeholder engagement activities, including the design and deployment of the FAITH anonymous survey. Section 4 explains the co-design workshops conducted to refine AI trustworthiness requirements. |
| **Task 2.3:** *"These considerations will be synthesized into AI trustworthiness requirements & metrics, plus best practice primarily targeted at data scientists and technology developers."* | **Section 4.5** summarizes the outcomes of the co-design workshops, including synthesized metrics, stakeholder-specific checklists, and refined requirements aimed at guiding data scientists and technology developers. |

## 1.3. Deliverable structure

Deliverable D2.5 is structured into five main sections and a comprehensive appendix. The structure is designed to enable the integration of insights, legal mappings, stakeholder engagement results, and co-creation outputs to support the continuous refinement of the FAITH AI Trustworthiness Assessment Framework (FAITH AI_TAF). The structure of D2.5 is presented below.

**Section 1** outlines the objectives and scope of the deliverable, its alignment with the Description of Action (DoA), and a brief description of how the document is organized. It sets the context for understanding how stakeholder engagement and legal mappings support the design and assessment of trustworthy AI systems in FAITH.

**Section 2** is divided into three key subsections with a brief overview of the conceptual framework derived from previous FAITH deliverables (notably D2.1). An in-depth mapping of the European and national legal frameworks applicable to each of the seven LSPs. This includes sector-specific legislation (e.g., GDPR, MDR, AI Act, NIS2, ePrivacy Directive) and domain-relevant provisions (e.g., maritime law, education law, media regulation, medical device certification). Each pilot is analyzed individually to identify applicable rules and their implications. The section concludes with an exploration of stakeholder support preferences and interaction expectations with AI systems, particularly focusing on dimensions such as explainability, autonomy, and user support mechanisms.

**Section 3** presents the methodology and results of the first wave of stakeholder engagement. It describes the multi-method approach used to identify and recruit stakeholders from all LSPs and beyond, ensuring ethical and inclusive engagement. The rationale and implementation of the FAITH anonymous survey are outlined, including how it was designed and deployed (via the GDPR-compliant EU Survey platform), and how the sample size was determined to ensure statistical robustness. The section further explains the data analysis procedures, integrating both quantitative and qualitative methods to extract meaningful insights. The results are then discussed, with specific focus on stakeholders' understanding of AI systems and their acceptance of AI in their respective domains, including comfort levels, perceived benefits, and sources of resistance.

**Section 4** focuses on the co-design workshops organized to refine the trustworthiness requirements for the FAITH AI_TAF. It explains the participatory methodology adopted, the iterative process for incorporating stakeholder feedback, and the output generated. These include revised trustworthiness metrics, stakeholder-specific checklists, and enhanced requirements tailored to the realities of each pilot domain. This section emphasizes how collaborative design contributes to the contextual relevance and usability of the FAITH AI_TAF.

**Section 5** offers a summary of conclusions and outlines the next steps. It reflects on the main findings of the deliverable and sets the stage for their integration into future work packages and deliverables, including those that will finalize the FAITH AI_TAF and assess its practical application. The appendix provides supplementary material such as the survey instrument, detailed analysis results, and completed templates from the FAITH Knowledge Base enhancement process per LSP.

The **Appendix** provides critical supplementary material to enhance the transparency, reproducibility, and utility of the results presented in the main body of the deliverable. It includes the complete survey instrument used for stakeholder engagement, enabling future reference and replication. Detailed Bayesian and statistical analysis results are also provided, supporting the interpretation of stakeholders' understanding and acceptance of AI. Furthermore, the appendix contains the completed templates from the FAITH Knowledge Base (KB) enhancement process, offering structured documentation of pilot-specific trustworthiness considerations. These annexed resources collectively contribute to the methodological depth of the deliverable and ensure that the FAITH approach remains grounded in empirical evidence and stakeholder realities.

# 2. Trustworthy and human-acceptable AI requirements

## 2.1. Key considerations and criteria for AI trustworthiness

Ensuring trust in Artificial Intelligence (AI)-based systems is a core objective of the FAITH project. Trustworthy AI is critical not only for compliance with emerging regulatory frameworks, such as the EU AI Act, but also for achieving human acceptance, ethical alignment, and stakeholder engagement across varied socio-technical environments. The FAITH Trustworthiness Assessment Framework (FAITH AI_TAF) offers a structured, risk-based methodology that enables the identification, assessment, and mitigation of threats to trust throughout the AI lifecycle. This approach is adapted to the specific needs of each Large-Scale Pilot (LSP) and is built upon a foundation of both technical and human-centric considerations.

According to the analysis presented in D2.1, trustworthiness in AI is inherently multidimensional. The FAITH AI_TAF explores dimensions derived from European and international best practices and standards, including guidance from the EU High-Level Expert Group on AI (HLEG), ENISA, NIST's AI Risk Management Framework, and ISO/IEC initiatives. Together, they form a comprehensive framework for evaluating and promoting trustworthy AI. To operationalize these dimensions, the FAITH AI_TAF adopts a six-phase, risk-based methodology. This begins with a cartography phase that maps the system's scope, components, stakeholders, and data flows. In the second phase, threats to trustworthiness are identified across technical, organizational, and social vectors. The third phase involves assessing the potential consequences of these threats, followed by an analysis of system vulnerabilities in phase four. In the fifth phase, risks are calculated by evaluating the likelihood and impact of potential failures. Finally, phase six provides a structured approach for selecting and applying countermeasures, including technical safeguards, governance mechanisms, and behavioural or social interventions. This end-to-end methodology reflects the complexity and interconnectedness of trust-related risks and ensures that assessments are tailored to the characteristics of the system and its context of use.

The FAITH AI_TAF is supported by three dedicated tools: (i) the FAITH AI TrustGuard, a checklist-based tool for conducting system-level risk assessments, (ii) the FAITH AI TrustSense, which profiles the psychosocial and trust-relevant attributes of AI participants; and (iii) the FAITH AI Model Hub, a metadata repository incorporating AI and data "passports" to improve traceability, documentation, and accountability. These tools provide both structure and flexibility, enabling LSP teams to operationalize trustworthiness principles while responding to specific legal, technical, and social requirements. A unique strength of the FAITH AI_TAF lies in its attention to human and organizational factors. D2.1 highlights how individual traits, group dynamics, and organizational maturity can directly influence the perceived and actual trustworthiness of AI systems. Human vulnerabilities such as cognitive biases, decision-making heuristics, or misaligned incentives can affect system design, deployment, and oversight. To address these, FAITH incorporates participant profiling and recommends mitigation strategies, including awareness training, co-design workshops, and behavioural nudges. This sociotechnical perspective reinforces the view that trust in AI is not only a matter of code and compliance, but also of communication, engagement, and shared understanding. Moreover, the FAITH AI_TAF is designed to align with the EU AI Act's risk classification model. Depending on the nature and domain of application, AI systems are categorised as unacceptable, high, limited, or minimal risk, with corresponding regulatory obligations. Many of the AI systems in FAITH's pilot domains—such as education, media, and healthcare—fall under the high-risk category, requiring stringent measures for explainability, bias control, human oversight, and documentation. The FAITH AI_TAF responds to these legal expectations by embedding them directly into its assessment phases and toolset. In summary, the FAITH AI_TAF offers a multidimensional, risk-oriented, and human-

centred approach to evaluating and promoting AI trustworthiness. By integrating regulatory compliance, technical robustness, and ethical integrity with practical assessment tools, the framework empowers LSPs and other stakeholders to develop and deploy AI solutions that are not only legally sound but also socially acceptable and aligned with European values.

A Knowledge Base (KB) is essential to support the FAITH AI_TAF because it provides a structured knowledge repository of threats, vulnerabilities, control measures, AI assets, AI lifecycle phases, and trustworthy AI attributes/dimensions (TWA). The FAITH KB is a relational database designed to support the FAITH AI Trustworthiness Assessment Framework (AI_TAF). Its functional purpose is to support the FAITH TrustGuard. The initial version is built on ENISA guidelines[1] which will be further enhanced and refined based on practical insights iteratively gained from the FAITH LSPs and other valuable sources. It helps users to identify and manage trust risks in AI-based systems by linking each challenge to specific control measures and recommendations. It also facilitates cross-sector learning by allowing stakeholders to apply lessons learned in one domain to similar problems in other domains, thus strengthening trust and resilience in AI development and deployment. The backbone of the FAITH KB starts from and expands the seven Trustworthiness Attributes (TWAs) as defined by the U.S. National Institute of Standards and Technology (NIST): (i) Valid & Reliable, (ii) Safe, (iii) Secure & Resilient, (iv) Accountable & Transparent, (vi) Explainable & Interpretable, (vii) Privacy-enhanced, and (viii) Fair - with harmful bias managed. These TWAs were selected for their coverage of both technical and human-centric aspects of trust, as well as their strong alignment with European regulatory principles, including those embedded in the EU AI Act and relevant ethical guidelines. Structurally, the FAITH KB integrates ENISA's AI assets, including: (i) processes, (ii) environments/tools, (iii) artefacts, (iv) models, (v) data, to enable precise threat identification and traceability. To contextualize these assessments temporally, the FAITH KB is organized around three distinct phases of the AI lifecycle as defined by the project: (i) design, (ii) testing, and (iii) production. This lifecycle staging ensures that trustworthiness can be assessed and reinforced at critical junctures, from early ideation to real-world deployment. Through the combination of the NIST TWAs, ENISA's assets classification, and a phased lifecycle approach, the FAITH KB offers a coherent, structured, and dynamic basis for capturing, analyzing, and co-developing trustworthiness insights across all seven LSPs. This foundation enables the project to move from abstract principles to context-aware, actionable assessments, as detailed in Section 4. More details regarding the structure and the content of the FAITH KB, which will be continuously updated through the project's lifecycle, will be presented in D2.3 "FAITH System Trust Modeller (STM) v1".

## 2.2. Applicable Legal Regimes per LSP

FAITH manages all Large-Scale Pilots (LSPs) in accordance with the project's Data Management Plan (DMP), as outlined in Deliverable D1.3. This ensures that each pilot adheres to established principles of responsible data governance, ethical compliance, and regulatory alignment. The DMP provides a detailed description of all LSPs and serves as a foundational reference for monitoring data handling practices, including data protection, privacy, and transparency across the project's lifecycle. To support the regulatory mapping that follows in this section, Table 2 offers a high-level overview of the seven Large-Scale Pilots (LSPs) implemented within the FAITH project. Each LSP addresses a specific domain of societal relevance—such as media, transport, education, robotics, environmental management, healthcare, and active ageing—and explores how AI technologies are applied in real-world scenarios. The table summarizes the focus and context of each pilot, along with the main partners involved. This summary is intended to provide the reviewer with a clear understanding of the

[1] https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms.

operational and technical environment of each LSP before delving into their associated legal and ethical frameworks.

**Table 2.** A summary of the FAITH LSPs.

| LSP No | LSP domain | Description | Involved Partners |
|---|---|---|---|
| 1 | Media | AI-based assistance to characterize disinformation in the news and claims | ATC, FH |
| 2 | Transportation | Use of AI for safer and sustainable public transit | CNR, MERMEC |
| 3 | Education | AI tools to constantly monitor student progress, to provide targeted feedback and assess the student's mastery | ICCS, EA |
| 4 | Robotics/Drones | AI driven maintenance of port infrastructure on data from underwater drones | VFP, APRA |
| 5 | Wastewater management | Hybrid AI models for wastewater treatment | SINTEF, Veas |
| 6 | Healthcare | Use of AI for the reporting of medical images in radiology workflows | FORTH, UNIPI, PAGNH, CNR |
| 7 | Active ageing | Use of AI for Active Ageing solutions | BRIDG, AOA |

### 2.2.1. Applicable Legal Regimes in LSP1 (Related Deliverable: D4.1)

AI Intelligent Coaching is a web-based application that refers to professional journalists. AI Intelligent Coaching consists of four main functionalities: 1) matching to debunked claims, 2) matching claims to other articles, 3) logical fallacy detection, and 4) hate speech detection.

The Data Protection Officer (DPO) of ATC has confirmed the Project Coordinator in writing that *"after a thorough examination of the Media LSP pilot plan and objectives, we approve the pilot's conduction by the ATC team with the participation of external professionals within the FAITH project framework"*.

The collected user data will be used only for the purposes of the project. In particular, the web application informs the user about the collected data and their use with the following text (which is subject to change based on the requirements of the project as the project evolves):

"The AI Intelligent Coaching Service is part of the FAITH EU project (GA #101135932), which aims to develop and validate a human-centric trustworthiness optimization ecosystem, which enables measuring, optimizing and counteracting the risks associated with AI adoption and trustworthiness in critical domains, namely robotics, education, media, transport, healthcare, active ageing, and industrial processes through seven international Large Scale Pilots. The AI Intelligent Coaching service is used for the media large scale pilot and aims to assist professional journalists to assess the credibility of the articles they write or read.

**Terms and Conditions**

Pursuant to art. 13 of EU Regulation No. 2016/679 (GDPR), and in general in observance of the principle of transparency set forth in the above Regulation, are herewith the following information regarding the processing of personal data of the users of the AI Intelligent Coaching Service.

**Purpose of data processing**

User data will be collected and processed for the following purposes:

a. so the user can retrieve their password if needed;

b. compliance with contractual requirements (KPIs) of the media large scale pilot for the FAITH project, i.e. to ensure that the AI System is being used by the required number of participants across all levels of professional experience.

**Type of data processed**

The following data categories may be processed for the above mentioned purposes:

a. personal particulars, i.e., e-mail address, password (encrypted), region, years of professional experience, experience in assessing other AI systems with respect to trustworthiness;

b. assessed articles.

**Processing method**

Your data shall be processed:

- manually and through automated means, on paper and/or electronic records;
- by individuals authorized and properly trained by ATC to perform such tasks;
- by using proper measures to ensure confidentiality and avoid access by non-authorized third parties;
- in contexts that do not compromise the personal dignity and the decorum of the person concerned, ensuring the necessary precautions to guarantee the protection in the use.

**Processing length**

The personal data collected shall be retained until three years after the end of the FAITH project.

**Data sharing**

All your data will be neither stored nor transmitted to third parties and only anonymized qualitative data will be used for the purposes of the FAITH project.

**Rights of the person concerned**

Pursuant to Section III of the GDPR, participants shall be entitled to exercise the following rights to:

- request access to own personal data and to obtain a copy thereof;
- if the legal conditions exist, oppose the processing of data, request their cancellation or exercise the right of limitation;
- withdraw at any time any consent given to the processing of personal data and such withdrawal will result in the end of treatment;
- if they are deemed as inaccurate or incomplete, request their correction or integration respectively;
- ask for the portability of own personal data to another subject in such a format that they can be easily consulted and used;
- advance a dispute to the Supervisory Authority (Privacy Authority);
- withdraw from the study at any time according to Art. 7 III 1 GDPR, as participation is voluntary.

**For specific information and contacts regarding the protection of personal data, including the exercise of the rights referred above, please send an e-mail to privacy@atc.gr.**

**Data Controller**

The Data Controller is ATC, with registered office in Rizariou, 10, Athens 152 33, Greece.

**Data Protection Officer**

The Data Protection Officer (DPO), designated by the Controller pursuant to art. 37 of the GDPR, may be reached via email: privacy@atc.gr.

*While using the AI Intelligent Coaching website, you hereby agree to the aforementioned terms and conditions.*

For any further questions please contact privacy@atc.gr"

Also, the user is notified about the limitations of the AI Intelligent Coaching service with the following disclaimer:

***Disclaimer***

*This AI-system provides information based on available sources and verification methods. While we strive for accuracy, we do not guarantee the completeness or correctness of the information provided. Users are responsible for interpreting and verifying the results before making decisions or taking action. We assume no liability for any consequences arising from the use of this system.*

According to AI Act, Annex III, AI Intelligent Coaching is not a system of high risk.

Furthermore, Article 50 of the AI Act imposes transparency obligations on AI systems interacting with humans. This includes informing users they are interacting with AI;

To comply with other ethical requirements, developers should also:

- Clearly explain the system's capabilities and limitations;
- Avoid deceptive design or outputs that could mislead users.

AI Intelligent Coaching must clearly communicate that outputs such as "fallacy detected" or "matched to debunked claim" are AI-generated suggestions, not factual determinations.

### 2.2.2. Applicable Legal Regimes in LSP2 (Related Deliverable: D5.1)

The following table (Table 3) provides an overview of the legislative and technical framework relevant to the deployment of AI systems in transport-related domains, especially within the Italian context. It highlights the intersection between data protection requirements—as defined by EU and national legislation—and sector-specific standards essential for the development, deployment, and certification of AI-enabled railway systems. The inclusion of GDPR and DPIA guidance reflects the importance of safeguarding personal data in high-risk AI applications, such as biometric recognition or smart city infrastructure. Meanwhile, references to CENELEC and CEI standards emphasize the strict operational and safety requirements imposed on AI technologies in regulated environments like railway systems. Together, these provisions shape the legal and operational landscape for trustworthy and compliant AI in critical infrastructure.

**Table 3.** Relevant legal provisions and technical standards for LSP2.

| Regulation/Provision | Scope |
|---|---|
| GDPR[2] | Personal data protection |
| Italian Privacy Code (Legislative Decree 196/2003, as amended)[3] | Personal data protection |
| EDPB Guidelines 07/2020 – Controllers and Processors[4] | Definition of Controller, Processor, Joint Controllers, and their relationships |

[2] Regulation (EU) No. 679/2016 (GDPR).
[3] Legislative Decree No. 196/2003 and amendments (Personal Data Protection Code).
[4] EDPB Guidelines "Guidelines 07/2020 on the concepts of controller and processor under the GDPR".

| | |
|---|---|
| WP29 Guidelines – DPIA and High-Risk Processing (WP248 Rev1)[5] | High-risk data processing and mandatory Data Protection Impact Assessments |
| Italian DPA Decision – DPIA List (Oct. 11, 2018)[6] | High-risk data processing and mandatory Data Protection Impact Assessments |
| WP29 Opinion 05/2014 – Anonymization Techniques[7] | Personal data anonymization techniques |
| Italian DPA Injunction – Smart City Project (Jan. 11, 2024)[8] | Penalty imposed on a municipality for AI-based "Smart City" projects |
| Italian DPA Guidelines – Biometrics & Graphometric Signature (2014)[9] | Biometric data processing |
| Italian Supreme Court Ruling No. 12967 (May 13, 2024)[10] | Biometric data processing |
| CENELEC EN 50128 – Railway Control Software[11] | Standard provisions for software used in railway control and protection systems |
| CENELEC EN 50657 – Rolling Stock Software[12] | Standard provisions for software used on board of rolling stock |
| CENELEC EN 50716 – Software Development Requirements[13] | Standard provisions for software used in railway applications |
| CEI EN 50155 – Environmental Standards for Electronic Devices[14] | Certification (obtained) |
| CEI EN 45545-2 – Fire Protection: Risk Class HL2[15] | Certification (obtained) |

### 2.2.3. Applicable Legal Regimes in LSP3 (Related Deliverable: D6.1)

This section describes the legal regulations that are relevant to the development and operation of artificial intelligence systems in education and the Educational LSP AI solution, at both European and national levels. The European Union has established a comprehensive legal framework for artificial intelligence (AI), with the Artificial Intelligence Act. The AI Act aims to ensure the safety and reliability of AI systems, as well as to promote innovation and the development of AI technology in the EU. The AI Act follows a risk-based approach to AI systems, classifying them into different risk levels, as shown in Figure 1.

[5] Working Party 29 Guidelines "Guidelines on Data Protection Impact Assessment and determining whether processing is likely to result in a high risk - WP248 Rev1".
[6] Decision of the Italian Data Protection Authority "List of processing operations subject to DPIA requirements under Article 35(4) of Regulation (EU) No. 2016/679 - October 11, 2018".
[7] Working Party 29 "Opinion 05/2014 on anonymization techniques".
[8] Decision of January 11, 2024 - web doc. 9977020 (Injunction Order of the Italian Data Protection Authority).
[9] Guidelines of the Italian Data Protection Authority "Guidelines on biometric recognition and graphometric signature, November 12, 2014".
[10] Italian Supreme Court, Civil Section I, ruling no. 12967 of May 13, 2024.
[11] CENELEC EN 50128: Railway Applications - Communication, Signaling, and Processing Systems - Software for Railway Control and Protection Systems.
[12] CENELEC EN 50657: Railway Applications - Rolling Stock Applications - Software on Board Rolling Stock.
[13] CENELEC EN 50716: Railway Applications - Requirements for Software Development.
[14] CEI EN 50155 for electronic devices: Vibration – Humidity – Temperature – EMC.
[15] CEI EN45545 Fire: Risk class HL2 EN 45545-2.

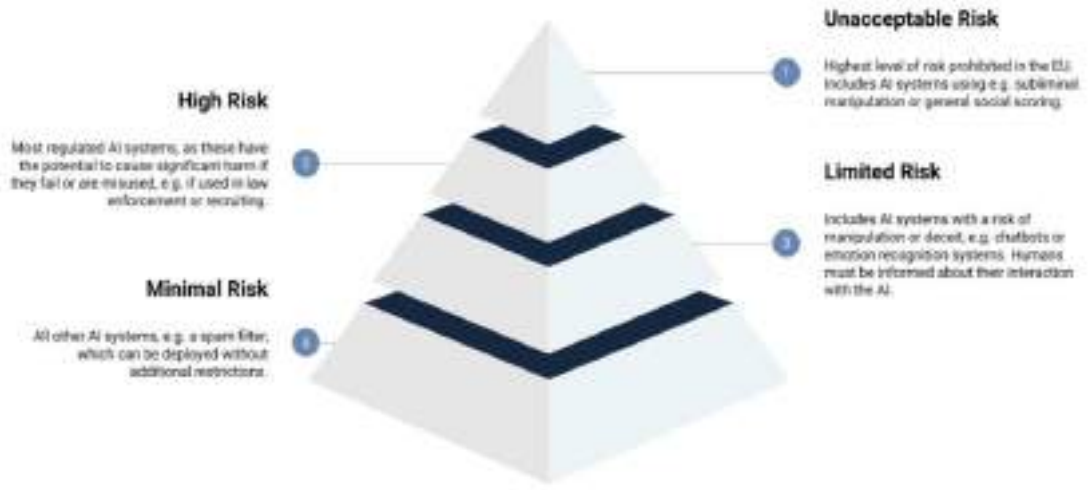



**Figure 1.** AI Risk Classification[16].

The EU AI Act adopts a risk-based approach. The risk levels based on the AI Act are as follows:

- Unacceptable Risk: AI systems that are considered a clear threat to the safety, livelihoods, and rights of people are prohibited.
- High Risk: AI systems that pose a significant risk to the health, safety, or fundamental rights of people are subject to strict requirements. This includes AI used in education, which can determine access to education and professional careers.
- Limited Risk: AI systems with specific transparency risks are subject to basic transparency requirements. For example, users must be informed when they are interacting with a chatbot.
- Minimal: AI systems that pose minimal or no risk can be used with few restrictions.

Since the Educational LSP AI solution is meant for education, the AI Act considers it a high-risk system. Though the platform itself does not keep any personal information, it's important to check that any AI used for learning or teaching follows the rules from the AI Act.

For high-risk AI systems, there are certain tough requirements that must be met.

- Risk assessment and mitigation: It is important for providers to evaluate and control the risks involved in their AI system.
- Data quality: Training the AI system with high-quality data can avoid the possibility of bias and discrimination.
- Technical documentation: Information about the AI system and what it is being used for should be shared with the authorities to help with assessing its compliance.
- Transparency: People should know when an AI system is working and how it functions.
- Human oversight: Appropriate systems must be put in place to supervise AI systems that carry important risks.

Laws in Greece are beginning to put in place to drive AI advancements while guaranteeing AI systems are used in an ethical and responsible way. Law 4961/2022 in Greece, "Emerging Information and Communication Technologies, Enhancing Digital Governance and Other Provisions," lays out how AI systems should be developed, marketed, and used. The law also outlines the accepted ways for public sector bodies, such as the majority of educational institutions, to make use of AI systems.

[16] https://www.trail-ml.com/blog/eu-ai-act-how-risk-is-classified.

### 2.2.4. Applicable Legal Regimes in LSP4 (Related Deliverable: D7.1)

#### *2.2.4.1. Introduction*

LSP4 leverages AI-driven Autonomous Underwater Vehicles (AUVs) for infrastructure monitoring in the ports of Ravenna and Valencia. This summary presents the current and upcoming legal and regulatory landscape, with distinct implications for the two AI systems under test (ATR and AC), and compliance strategies to ensure trustworthiness and alignment with European and national legislation. The focus is on demonstrating full legal compliance. Although no personal or human-related data is processed (as operations occur fully underwater), all relevant legislative areas have been considered.

#### *2.2.4.2. Legal overview*

The following table (Table 4) provides a comparative overview of the current regulatory frameworks applicable to key domains relevant to LSP4 (Port Infrastructure), focusing on EU-level, Spanish, and Italian legislation. It highlights domain-specific obligations and offers justifications for their applicability—or lack thereof—in the operational context of the pilot.

**Table 4.** Mapping of current EU and National legal frameworks relevant to LSP4 operations.

| Domain | Current AU Legislation | Spain | Italy | LSP Applicability / Justification |
|---|---|---|---|---|
| Trustworthy AI | GDPR (EU 2016/679), Ethics Guidelines for Trustworthy AI[17,18] | LOPDGDD (Organic Law 3/2018)[19] | Decreto Legislativo 101/2018[20] | **Not applicable to LSP4** – No personal or biometric data is processed underwater. No human data is captured or analyzed. |
| Maritime & Port Operations | EU Port Services Regulation (Reg. 2017/352)[21], IMO COLREGs[22] | Puertos del Estado (State Ports) framework law[23] | Codice della Navigazione, Decreto Legislativo 169/2016[24] | **Applicable** – AUV missions are coordinated with local Maritime and Port Authorities; permissions are secured before operations.<br>UAVs (AUVs) are subject to temporary operational guidelines defined by the ports and comply with safety navigation protocols. |
| Use of Unmanned Vehicles (UAVs) | Directive 2009/45/EC[25]; EMSA Guidelines for Remotely Operated Vehicles[26] | Spanish Maritime Navigation Law[27] | Codice della Navigazione (Art. 123-129)[24] | **Applicable** – As AUVs operate in underwater port areas, not the open sea, they are governed by port-specific permissions and guidelines. |

[17] https://eur-lex.europa.eu/eli/reg/2016/679/oj
[18] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[19] https://www.boe.es/buscar/doc.php?id=BOE-A-2018-16673
[20] https://www.gazzettaufficiale.it/eli/id/2018/09/04/18G00129/sg
[21] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32017R0352
[22] https://www.imo.org/en/About/Conventions/Pages/COLREG.aspx
[23] https://www.boe.es/buscar/doc.php?id=BOE-A-2014-7876
[24] https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2016-08-04;169
[25] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32009L0045
[26] https://emsa.europa.eu/mass.html
[27] https://www.boe.es/buscar/act.php?id=BOE-A-2014-7876

| | | | | |
|---|---|---|---|---|
| Environmental Protection | Marine Strategy Framework Directive (MSFD, 2008/56/EC)[28] ; Habitats Directive (92/43/EEC)[29] | PERS-certified local rules in Valenciaport [30] | ARPA regional environmental policies[31] | **Applicable with limitations** – MSFD not directly enforced inside ports, but operations comply with local PERS and ARPA environmental safeguards. |

*2.2.4.3. Future legal developments*

Table 5 outlines key upcoming EU legislative instruments that may impact the operation of AI-based systems in LSP4. It highlights expected timelines, relevance to specific AI components, potential implications, and current preparedness strategies adopted within the FAITH framework.

**Table 5.** Forthcoming regulatory developments and strategic preparedness for LSP4.

| Upcoming Legislation | Expected Entry | Relevant AI Systems | Implication for LSP4 | Preparedness / Strategy |
|---|---|---|---|---|
| EU Artificial Intelligence Act[32] | 2025–2026 | ATR, AC | Likely to be classified as "high-risk" AI due to use in critical infrastructure. Requires CE marking, transparency, monitoring. | FAITH pilots to apply voluntary alignment: operator oversight, explainability, continuous logging. |
| Cyber Resilience Act (CRA)[33] | 2025–2026 | ATR, AC | Will require secure software, update mechanisms, and incident reporting. | Modular system design allows future compliance with secure development and update pipelines. |
| Revised MSFD[34] | TBD (post-2025) | AC | Potential expansion of marine habitat protection could affect sonar use. | Missions avoid sensitive habitats; port compliance already integrated (ARPA, PERS). |

*2.2.4.4. AI System 1 - Automatic target recognition (ATR)*

**Current Legal Considerations:**

- **Data Protection**: **Not applicable** to ATR as no personal data is processed. The system solely uses sonar and optical imagery of underwater structures and objects.
- **Liability & Accountability**: Misclassification of hazards is a risk. ATR outputs are logged and supervised in real time to ensure operational traceability and human oversight, as required under current civil liability frameworks.
- **Transparency and Explainability**: Existing obligations under Trustworthy AI Guidelines are voluntarily followed.
- **Operator Supervision**: Full human-on-the-loop control is maintained, in line with good practice and voluntary alignment with AI Trust Guidelines.

**Compliance Strategies:**

- All detections are logged and traceable.
- No automated critical decisions are made without operator validation.

[28] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32008L0056
[29] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:31992L0043
[30] https://www.ecoports.com/pers
[31] https://www.arpae.it/it/temi-ambientali/mare
[32] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206
[33] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0454
[34] https://www.europarl.europa.eu/legislative-train/theme-a-european-green-deal/file-review-of-the-marine-strategy-framework-directive

- Explainability tools will be embedded to support human understanding.

**Impacts:**

- Fully compliant with current regulations.
- Supports risk mitigation for port infrastructure monitoring.
- Compliant with safety and environmental constraints within port areas.

#### 2.4.4.5. AI System 2 - Autonomous coverage (AC)

**Current Legal Considerations:**

- **Data Protection**: **Not applicable** to AC as no personal data is processed.
- **Maritime Navigation Safety**: As AC computes optimal paths for underwater mapping, compliance with port authority regulations is ensured through pre-approved mission planning.
- **Environmental Protections**: Though the MSFD is not directly applied to port basins, multibeam echosounders are operated under ARPA and PERS environmental protocols to avoid disturbing marine life.
- **Operator Supervision**: Full human-on-the-loop control is maintained, in line with good practice and voluntary alignment with AI Trust Guidelines.

**Compliance Strategies:**

- Mission zones, exclusion areas, and traffic overlays are integrated into planning software.
- Sonar equipment is pre-authorized and calibrated to minimize environmental impact.

**Impacts:**

- Compliant with safety and environmental constraints within port areas.
- Enhances survey efficiency with negligible risk to infrastructure or operations.

LSP4 complies with all current EU, Spanish, and Italian regulations. Although no personal or human data is processed, the LSP ensures transparency, safety, and environmental respect. As AI systems operate underwater in restricted, pre-authorized port areas, human oversight, data traceability, and regulatory engagement are ensured at all phases. Future legislation such as the AI Act and CRA are proactively addressed through existing design choices and documentation practices. FAITH's LSP4 pilot provides a legally sound foundation for trustworthy AI in maritime robotics. A legal monitoring strategy will ensure readiness for evolving regulatory landscapes ahead of deployment scale-up.

### 2.2.5. Applicable Legal Regimes in LSP5 (Related Deliverable: D8.1)

Wastewater treatment is a complex process covered by a broad range of relevant legal regimes. In the following, we summarize some of them related to (a) the application of AI, (b) wastewater treatment, and (c) work environment. We first provide an overview of relevant European legislation, focusing on (a) and (b), before detailing the relevant national legislation for Norway, focusing on (b) and (c). We also provide a summary of relevant quality policies and certifications. It should be noted that the summarized legislation is far from comprehensive, given the broad range of relevant legal regimes for wastewater treatment.

#### *2.2.5.1. Relevant European legislation*

Of the relevant European legislation, we in particular mention two: The EU AI Act and the Urban Wastewater Treatment Directive.

**EU AI Act**

The AI Act establishes a comprehensive framework for the development, deployment, and use of AI systems within the EU. The implications of the EU AI Act for the use of AI prediction models for wastewater treatment will depend on the models are considered high-risk applications or not. Consideration as high-risk applications poses important requirements for, e.g., risk management, transparency, and accountability. Norway, as a member of the EEA, is in the process of implementing the AI Act as legislation, and the government aims for the legislation to be in force during the latter part of 2026.

**Urban Wastewater Treatment Directive (UWWTD)**

The UWWTD aims to protect human health and the environment from harms following from untreated urban wastewater. It entails the obligation in EU countries to collect and treat wastewater in urban areas, including the removal of nutrients and micropollutants, and monitor wastewater with regards to health parameters. Norway, as member of the EEA, has implemented UWWTD in national law – as part of the national pollution regulations. A revised version of the UWWTD, which is in force in the EU from 2025, is in the process of being implemented as legislation in Norway.

#### *2.2.5.2. Relevant Norwegian Legislation*

Of the relevant Norwegian legislation, we in particular mention legislation on pollution and waste control, as well as legislation on work environment. Both detailed below. We also mention assessments conducted to ensure regulatory compliance. Legislation on AI, following from the implementation of the AI Act in Norwegian law, will also be part of relevant Norwegian legislation. However, any specific national details, if relevant, are not established yet as the implementation is still in process. Hence, we do not address this specifically under the section on Norwegian legislation.

**Norwegian legislation on pollution and waste control**

Pollution and waste control is regulated by national legislation, including the following:

- **Pollution Control Act ('Forurensningsloven').** This legislation aims to protect the environment against pollution and promote better waste management. It includes provisions on the prevention and limitation of pollution and waste management.
- Pollution Regulations ('Forurensningsforskriften'). These regulations address pollution control measures, including cleanup of contaminated soil, noise mapping, and limits, and discharge to public sewage systems. The regulation also specifies requirements on wastewater treatment, in terms of types of treatment and level of cleaning.
- Waste Regulations ('Avfallsforskriften'). These regulations cover various aspects of waste management, including the sorting and recycling of waste types and handling of hazardous waste.
- Nature Diversity Act ('Naturmangfoldloven'). This Act focuses on the conservation of biodiversity and sustainable use of natural resources. It includes provisions on prevention of environmental damage.

The national legislation on pollution, as listed above, also serves to implement and adapt the European UWWTD as Norwegian law.

**Legislation on work environment**

The work environment in an industrial environment like for wastewater treatment is subject to a range of legislation. Some key legislation includes:

- Working Environment Act ('Arbeidsmiljøloven'). This Act ensures safe and healthy working conditions, including requirements for risk assessment, training, and use of personal protective equipment 6.

- Internal Control Regulations ('Internkontrollforskriften'). These regulations mandate systematic health, environment, and safety work within organizations. They include requirements for documentation, risk assessment, and continuous improvement.

**Compliance assessments**

Assessments of compliance are conducted to ensure that operational procedures adhere to relevant legal frameworks, permits and stakeholder interests. This includes, e.g., assessment of compliance with the legislation listed above, discharge permits for wastewater, and to relevant stakeholders.

*2.2.5.3. Certification standards*

Certification standards are valuable for documenting the quality needed in operations and enable continued improvement in wastewater treatment. The following standards are particularly relevant for LSP5:

**Quality management system - ISO 9001**

ISO 9001 is an international standard for quality management systems. The standard addresses quality of performance through a process approach, with concern for, e.g., reliability , effectiveness, ability to adapt, and continuous improvement. Identification and mitigation of risks is also included as key to ISO 9001.

**Environmental management system – ISO 14001**

ISO 14001 is an international standard for environmental management systems. It supports the management of environmental responsibilities, minimize environmental impact, and ensure compliance with regulation of environmental relevance. As part of certification, both for ISO 9001 and ISO 14001, external and internal auditing is required at regular intervals decided in accordance with the needs of the organization and requirements in the standard.

### 2.2.6. Applicable Legal Regimes in LSP6 (Related Deliverable: D9.1)

The healthcare sector is undergoing a rapid digital transformation, with technology becoming increasingly integrated into nearly every aspect of medical practice. Software, in particular, has emerged as a critical component of this evolution, playing an essential role in a wide array of medical applications and interventions. In some instances, software itself meets the definition of a medical device, highlighting its growing significance in the delivery of healthcare services. Medical software is proving to be an invaluable asset in modern healthcare, offering solutions that can streamline workflows, reduce operational costs, and optimize the allocation of time and resources for both healthcare professionals and patients. Software as a Medical Device (SaMD) is a term that has gained prominence in the healthcare technology sector, referring to software that is intended to be used for one or more medical purposes without being part of a hardware medical device. The key characteristics that define SaMD include its specific intended medical purpose, its ability to function independently of any medical hardware, and its capacity to operate on general-purpose computing platforms such as smartphones, tablets, and cloud servers. Examples of SaMD in the healthcare domain are diverse and include applications such as AI-powered tools for diagnostic imaging, mobile applications designed for managing chronic diseases by monitoring patient-generated data and providing treatment recommendations, and software platforms utilized for delivering digital therapeutics. In the LSP6 at the healthcare domain the AI-based SUT falls in the category of a SaMD (described in detail in the deliverable D9.1 "Report pilot coordination and experimental infrastructure – Healthcare LSP").

Moreover, in the development process of SaMD and specifically AI-based SaMD, it is very important

to follow Good Clinical Practices (GCP)[35] to protect the rights, safety, and health of patients and to make sure that clinical research is done in an ethical and scientific way. GCP gives researchers a way to plan, carry out, document, and report clinical studies that use SaMD, especially when the software helps with diagnosis, monitoring, or therapy decisions. It is the developers' job to make sure that clinical evaluations of SaMD follow ISO 14155:2020, which sets the rules for clinical studies of medical devices, even those that are software-based. This includes making a strong clinical evaluation plan, getting informed consent, keeping the data safe, and making sure that the evidence used to show clinical success can be traced. GCP compliance also means engaging with ethics committees and other authorities to make sure that the rules are followed. In the context of trustworthy AI, adopting GCP makes AI-enabled SaMD more accountable and open, which helps with risk management, repeatability, and post-market surveillance when regulatory frameworks evolve.

#### *2.2.6.1. Overview of the Medical Devices Regulation (MDR) (EU) 2017/745*

The European Union (EU) has established a comprehensive legal framework for medical devices through Regulation (EU) 2017/745 on medical devices (MDR), which became fully applicable on May 26, 2021. This regulation replaced the earlier Medical Devices Directive (MDD) and the Directive on Active Implantable Medical Devices (AIMDD). The primary objective of these new regulations is to ensure a high level of safety and performance for medical devices made available on the EU market, while also fostering innovation within the sector. The MDR aims to create a robust, transparent, and sustainable regulatory system that addresses the evolving needs and technological advancements in the field. To facilitate a smooth transition from the previous directives, the MDR included specific transition periods during which devices that were legally placed on the market under the old rules can continue to be made available, provided certain conditions are met. The duration of these transition periods varies depending on the risk classification of the device[36]. MDR introduces more rigorous procedures for the conformity assessment of medical devices, encompassing not only new products but also those that have been marketed in the EU for many years[37]. These enhanced procedures reflect the EU's commitment to strengthening the oversight of medical devices throughout their lifecycle, from design and manufacture to post-market surveillance.

**Definition and Qualification of Medical Device Software (MDSW) under EU MDR (MDCG 2019-11)**

The Medical Device Coordination Group (MDCG), a body established under the MDR to support its consistent implementation, has issued guidance document MDCG 2019-11 which provides critical insights into the qualification and classification of software under the MDR. This guidance defines Medical Device Software (MDSW) as software intended to be used, whether alone or in combination with hardware, for a medical purpose as defined in the MDR. This definition is intentionally broad, encompassing both standalone software applications (often referred to as Software as a Medical Device or SaMD in other contexts) and software that drives or influences the use of a hardware medical device, including embedded software[38]. For software to be qualified as MDSW, it must have a clearly defined intended medical purpose. This purpose can include a wide range of medical functions, such as diagnosis, prevention, monitoring, prediction, prognosis, treatment, or alleviation of disease, injury, or disability[38]. However, software that merely facilitates administrative tasks, such as data storage, archival, communication, or basic search functionalities, without performing a specific

[35] European Medicines Agency. (n.d.). Good clinical practice (GCP). Retrieved June 25, 2025, from European Medicines Agency website: https://www.ema.europa.eu/en/human-regulatory-overview/research-development/compliance-research-development/good-clinical-practice.

[36] SaMD – Software as a Medical Device: EU MDR Rule 11 Classification & CE Marking - Maven Profcon Services, accessed May 19, 2025, https://mavenprofserv.com/blog/ce-medical-device-software-samd/.

[37] Medicines and Medical Devices | Fact Sheets on the European Union, accessed May 19, 2025, https://www.europarl.europa.eu/factsheets/en/sheet/50/medicines-and-medical-devices.

[38] Software as medical device: definition and classification, accessed May 19, 2025, https://blog.johner-institute.com/iec-62304-medical-software/software-as-medical-device-definition-and-classification/.

medical function or adding medical value through analysis or interpretation, is generally not considered MDSW[39]. Conversely, software that incorporates a "call to action" directly related to the aspects covered by Article 2 of the MDR, such as an app that advises a patient to consult their doctor based on collected data, is likely to be classified as MDSW[39].

*2.2.6.2. Classification Rules for SaMD under EU MDR*

VIII of the MDR lays out the rules for classifying medical devices based on the level of risk associated with their intended use. Rule 11 within this annex is specifically dedicated to the classification of medical software, including SaMD. Under Rule 11a, software intended to provide information that is used to take decisions with diagnosis or therapeutic purposes is generally classified as Class IIa. However, this classification can be elevated based on the potential impact of those decisions. If such decisions may cause a serious deterioration of a person's state of health or necessitate a surgical intervention, the software is classified as Class IIb. In the most critical cases, where the decisions may cause death or an irreversible deterioration of a person's state of health, the software is classified as Class III. Rule 11b addresses software intended for monitoring physiological processes. Finally, all other medical device software that does not fall under the criteria of Rule 11a or 11b is classified as Class I under Rule 11c[40].

**Conformity Assessment Procedures and CE Marking for SaMD**

To be legally placed on the EU market, SaMD must undergo a conformity assessment procedure to demonstrate that it meets the safety and performance requirements stipulated in the MDR. Successful completion of this assessment allows the manufacturer to affix **the CE marking to the SaMD,** signifying its compliance with EU regulations[41]. The conformity assessment process typically involves an audit of the manufacturer's Quality Management System (QMS) based on the ISO standard ISO 13485 and a thorough review of the technical documentation related to the SaMD (Figure 2)[42].




**Figure 2.** Key steps to an effective quality management system (QMS).

For SaMD classified as Class IIa or higher risk, the assessment must be carried out by an EU-accredited Notified Body, which is an independent organization designated by an EU Member State to evaluate

[39] Software as a Medical Device (SaMD) | MDR Compliance Guide - Mantra Systems, accessed May 19, 2025, https://www.mantrasystems.co.uk/eu-mdr-compliance/software-as-a-medical-device-samd.
[40] health.ec.europa.eu, accessed May 19, 2025, https://health.ec.europa.eu/system/files/2020-09/md_mdcg_2019_11_guidance_qualification_classification_software_en_0.pdf
[41] Medical Device Regulation in Europe: A Comprehensive Guide - Masson International, accessed May 19, 2025, https://www.massoninternational.com/blog/medical-device-regulation-in-europe.
[42] Medical devices | European Medicines Agency (EMA), accessed May 19, 2025, https://www.ema.europa.eu/en/human-regulatory-overview/medical-devices.

the compliance of medical devices. Class I devices, however, can generally undergo a self-assessment procedure. According to Figure 3, the steps involved in the conformity assessment process usually include defining the intended purpose of the SaMD, accurately determining its risk classification, establishing and implementing a QMS that adheres to the MDR requirements, compiling comprehensive technical documentation as detailed in Annexes II and III of the MDR, and finally, undergoing the assessment by a Notified Body if the risk class necessitates it.

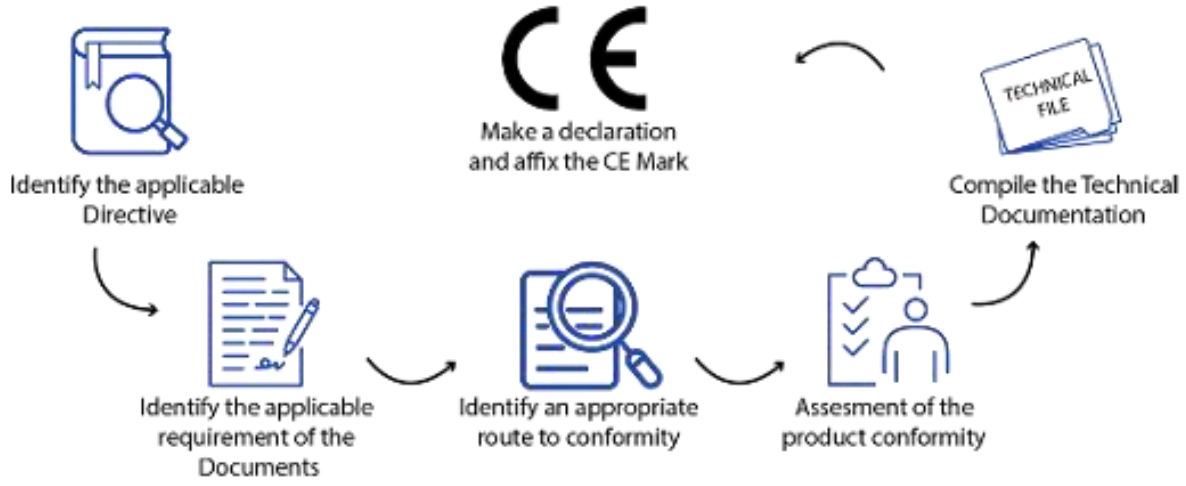



**Figure 3.** Steps for CE marking and conformity assessment.

## Clinical Evaluation Requirements for SaMD

The EU MDR places a strong emphasis on clinical evaluation as a continuous process that manufacturers must undertake to demonstrate the safety and performance of their medical devices, including SaMD, throughout their entire lifecycle. This process involves systematically generating, collecting, analyzing, and assessing the clinical data pertaining to the SaMD to ensure it is safe to use and perform as intended. Manufacturers are required to develop a detailed Clinical Evaluation Plan (CEP) that outlines the objectives, methodology, and timeline for the clinical evaluation process, including the types of clinical data to be used and the criteria for assessing the device's safety and performance. Following the execution of the plan, manufacturers must compile a comprehensive Clinical Evaluation Report (CER) that documents the entire clinical evaluation process, the clinical data analyzed, and the conclusions drawn regarding the safety and performance of the SaMD. The clinical evidence for SaMD should specifically address three key components: a valid clinical association, which demonstrates a link between the software's output and the targeted clinical condition; the technical performance of the software, including its accuracy, reliability, and usability; and the clinical performance of the software, which proves its ability to achieve its intended medical purpose and provide a tangible clinical benefit to the patient. In certain cases, particularly for lower-risk SaMD, manufacturers may be able to justify demonstrating conformity based solely on the results of non-clinical testing methods, such as performance evaluations, bench testing, preclinical evaluations, and usability assessments. However, even in these situations, a CER is still required to provide well-reasoned justification for the adequacy of the non-clinical data. Furthermore, post-market clinical follow-up (PMCF) is a crucial element of the clinical evaluation process, requiring manufacturers to actively and continuously monitor the safety and performance of their SaMD once it is on the market[43].

## Data Privacy and Security Requirements (GDPR)

[43] EU MDR: Definition, Timelines, Requirements, and Compliance - SimplerQMS, accessed May 19, 2025, https://simplerqms.com/eu-mdr/.

The General Data Protection Regulation (GDPR) (EU) 2016/679[44] establishes a comprehensive legal framework for the protection of personal data and applies to the processing of personal data of individuals within the European Union (EU), regardless of whether the data controller or processor is established within the EU or in a third country. Data concerning health is specifically recognized under the GDPR as a special category of personal data, warranting a higher level of protection due to its sensitive nature. Manufacturers of SaMD that process the personal data of EU residents are obligated to comply with the core principles of the GDPR. These principles include **lawfulness, fairness, and transparency in processing; purpose limitation, ensuring data is collected for specified, explicit, and legitimate purposes; data minimization, collecting only necessary data; accuracy, keeping data up to date and accurate; storage limitation, retaining data only as long as necessary; integrity and confidentiality, processing data securely; and accountability,** demonstrating compliance with the GDPR[44]. The processing of health data generally requires explicit consent from the data subject, although the GDPR provides exceptions for processing necessary for purposes such as preventive or occupational medicine, medical diagnosis, the provision of health or social care, or reasons of public interest in the area of public health[45]. The GDPR also grants individuals a range of rights concerning their personal data, including the right to be informed about the processing, the right to access their data, the right to rectification, the right to erasure (also known as the "right to be forgotten"), the right to restrict processing, and the right to data portability.[73] In the event of a personal data breach that is likely to result in a risk to the rights and freedoms of individuals, the data controller is required to notify the competent supervisory authority without undue delay. If the breach poses a high risk to the affected individuals, they must also be informed directly.

### Cybersecurity Requirements

The EU MDR explicitly mandates that medical devices incorporating electronic programmable systems, including software, must be designed to ensure their reliability, repeatability, and performance according to their intended use. Furthermore, manufacturers are required to implement appropriate measures to eliminate or minimize risks associated with single fault conditions. The Medical Device Coordination Group (MDCG) has issued guidance MDCG 2019-16 to provide manufacturers with recommendations on fulfilling the cybersecurity requirements outlined in Annex I of MDR. This guidance emphasizes the importance of adopting a risk-based approach to cybersecurity, integrating security considerations into the design phase ("security by design"), and ensuring ongoing vigilance and monitoring of medical devices once they are on the market.[38] Manufacturers are expected to conduct thorough cybersecurity risk assessments to identify potential vulnerabilities and threats, implement robust security controls to mitigate these risks, and establish clear processes for the transparent disclosure of vulnerabilities[46]. Additionally, MDCG 2023-4 offers specific guidance on the regulatory considerations for Medical Device Software (MDSW) that is intended to work in combination with hardware or hardware components. This guidance clarifies the responsibilities of manufacturers in ensuring the safety, performance, and interoperability of such combined systems[47].

### National Competent Authorities in EU Member States (e.g., Greece)

In the European Union, while the MDR provides a harmonized regulatory framework, the implementation and enforcement of these regulations are also overseen by National Competent Authorities (NCAs) within each Member State[48]. These authorities are responsible for transposing the

[44] General Data Protection Regulation (GDPR) - TrustArc, accessed May 19, 2025, https://trustarc.com/regulations/gdpr/.
[45] What does the EU GDPR mean for your medical device product? - Extra Horizon, accessed May 19, 2025, https://www.extrahorizon.com/gdpr-software-medtech-medical-device-eu-regulation-mdr-data-privacy.
[46] MDCG 2019-16 & MedTech Cybersecurity - Blue Goat Cyber, accessed May 19, 2025, https://bluegoatcyber.com/blog/mdcg-2019-16-and-medical-device-security.
[47] health.ec.europa.eu, accessed May 19, 2025, https://health.ec.europa.eu/system/files/2023-10/md_mdcg_2023-4_software_en.pdf.
[48] Competent Authority, Notified Body, Authorized Representative – who's who and what do they do? - Rimsys, accessed May 19, 2025, https://www.rimsys.io/blog/eu-competent-authority-notified-body-authorized-representative.

EU regulations into their national laws and for ensuring that medical devices placed on their markets comply with the requirements. Additionally, EU Member States designate accredited Notified Bodies that are authorized to conduct conformity assessments for medical devices. For instance, in Greece, the National Organization for Medicines (EOF) serves as the national competent authority responsible for regulating medical devices[49]. Manufacturers intending to market their SaMD in Greece must comply with the MDR and any specific national requirements set forth by the EOF. One important aspect to consider is that EU Member States may have their own language requirements for the labeling and instructions for use of medical devices made available in their territories[50].

**Note:** In Greece, the official language for information intended for patients or end-users is Greek, although exceptions may be made on a case-by-case basis for products intended exclusively for professional use, where English may be acceptable[50].

*2.2.6.3. AI Act in the Healthcare domain*

The EU has recently approved the EU AI Act, the world's first comprehensive legal framework for AI. This regulation, which entered into force on August 1, 2024, and will be fully applicable by August 2, 2026 (with specific timelines for certain provisions), has a significant impact on AI-based medical devices. In brief:

- AI systems intended to be used as a safety component of a product or that are themselves a product subject to the MDR or IVDR are classified as high-risk under the AI Act. This includes many AI-based SaMD applications used for diagnosis, monitoring, and treatment.
- Providers (manufacturers) of high-risk AI systems, including AI-enabled medical devices, must comply with several obligations, such as implementing risk-mitigation systems, using high-quality data sets, providing clear user information, ensuring human oversight, and establishing a quality management system.
- High-risk AI-based medtech products must undergo third-party conformity assessments to ensure compliance with both the AI Act and the MDR/IVDR.
- Transparency requirements mandate clear documentation about how the AI system makes decisions, ensuring outputs are explainable to both medical professionals and patients.
- Providers need to establish post-market monitoring procedures to track the performance of their AI systems after they are placed on the market.

To conclude, the following tables (Tables 6-8) describe in a nutshell the regulatory landscape of an AI-based SaMD in the EU.

**Table 6.** EU MDR software classification requirements relevant to LSP6.

| Feature | Decision Support (Diagnosis/Therapeutic) | Monitoring Physiological Processes | All Other Software |
|---|---|---|---|
| **EU MDR (Rule 11)** | Class IIa (General); Class IIb (Serious Deterioration/Surgery); Class III (Death/Irreversible Deterioration) | Class IIa (General); Class IIb (Vital Parameters, Immediate Danger) | Class I |

**Table 7.** Overview of core GDPR requirements relevant to LSP6.

| Feature | Primary Legislation | Scope | Consent | Data Breach Notification |
|---|---|---|---|---|

[49] MDR Language Requirements in Greece - AbroadLink Translations, accessed May 19, 2025, https://abroadlink.com/blog/mdr-language-requirements-greece.

[50] Medical Devices MDR - Language requirements for manufacturers - Public Health - European Union, accessed May 19, 2025, https://health.ec.europa.eu/document/download/aa9760e3-c864-4173-8b16-d790dac66d74_en?filename=md_sector_lang-req-table-mdr.pdf.

| | | | | |
|---|---|---|---|---|
| **European Union (GDPR)** | General Data Protection Regulation (EU) 2016/679 | Personal data of EU residents | Explicit consent is often required for health data | Supervisory authority within 72 hours; individuals if high risk |

**Table 8.** Key cybersecurity and risk management requirements relevant to LSP6 under EU MDR and MDCG.

| Feature | Key Principles | Guidance/Regulations | Key Requirements |
|---|---|---|---|
| **European Union (MDR & MDCG)** | Risk-based approach, security by design, ongoing vigilance, data protection by GDPR | MDR Annex I, MDCG 2019-16, MDCG 2023-4, ISO 27001 | Risk assessment & management, security features, minimum IT requirements, postmarket surveillance, vigilance reporting |

### 2.2.7. Applicable Legal Regimes in LSP7 (Related Deliverable: D10.1)

The Active Ageing LSP (LSP7) focuses on deploying non-intrusive, AI-powered home monitoring solutions to promote independent living for older adults and support caregivers in the region of Spain. The AI system under test (SUT) in this pilot is based on Vitana, an existing solution enhanced with an AI framework. The Vitana system processes personal data, including behavioral data derived from sensors (e.g., sleep patterns, mobility metrics) and potential feedback from interviews or questionnaires.

**European regulations**

- **The General Data Protection Regulation (GDPR)**[51] governs how personal data is collected, used, and stored, ensuring it is processed lawfully, transparently, and securely. The articles that apply include:
  - **Special-category data**: Processing health-related data for "*scientific research and statistical purposes*" is permitted (**Art 9(2)(j)**) if appropriate safeguards relating to data-minimization are implemented (**Art 89(1)**). Pseudonymisation, data minimisation, ethics committee approval and DPIA have been implemented.
  - **Consent and transparency**: Consent (**Art 4(11)**) can be given to allow the processing of special-category data (**Art 9(2)(a)**). LSP7 consent form provides all details about the purposes of the data processing and the data subject's rights that are clearly conveyed (**Art 12-13**).
  - **Individual rights**: The LSP7 Data Protection Policy (DPP) ensures the preservation of rights related to data access (**Art 15**), rectification (**Art 16**), erasure (**Art 17**), portability (**Art 20**), restriction (**Art 18**) and objection (**Art 21**).
  - **Data protection impact assessment (DPIA)**: In case of high risk to individuals' rights, which is the case of special-category data, conduct a DPIA, appoint a DPO, adhere to the privacy-by-design principle and maintain records of processing activities (**Art 25-35**).
- **The European Artificial Intelligence Act (EU AI Act)**[52] aims to ensure that AI is safe, trustworthy and respects fundamental rights by establishing the requirements for trustworthiness and risk management. The AI system implemented by LSP7 falls under the "high-risk" definition of the AI Act. Specifically, the articles that should be met are:
  - **Risk-management system** (requested for high-risk AI systems). The system will implement the FAITH AI Framework and measures such as continuous performance monitoring, robustness or bias (**Art 9**).

[51] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data [2016] OJ L119/1 (General Data Protection Regulation).
[52] Regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence [2024] (AI Act).

- **High Data Quality Standards** (**Art 10**): The system uses representative, bias-mitigated datasets and applies rigorous data governance practices. All data is pseudonymised, securely stored, and processed in accordance with GDPR.
- **Technical documentation** (**Art 11**): Documentation is available for sensor logs, system architecture, model training and validation protocols, and traceability tools.
- **Transparency Requirements** (**Art 13**): The system and its objective are clearly explained in the consent form. Additionally, it uses SHAP-based explainability and user-friendly, self-explanatory tools.
- **Human oversight** (**Art 14**): Human operators continuously monitor the system and intervene if necessary.

- **EU Charter of Fundamental Rights (EU Charter)**[53] goal is to protect fundamental rights in the EU, adapting to societal changes, progress, and technological developments. Specifically, LSP7 must respect the right to health (**Art 35**) and privacy (**Art 8**) and prevent discrimination (**Art 21**). This is ensured through early health risk detection, GDPR-compliant data practices, and fairness measures in AI to avoid bias and ensure equitable care.
- **ePrivacy Directive (2009/136/EC)**[54] focuses on the protection of privacy and personal data in the electronic communications sector. The LSP7 complies with this directive by ensuring informed consent for the use of home sensors, securing communications through encryption, and aligning data collection practices with privacy-by-design principles.

**Cybersecurity regulations**

These regulations impose obligations during the design, development, and deployment of the system. LSP7 follows secure-by-design principles, performs robustness testing periodically and has traceability features.

- **NIS2 Directive**[55]: applies to providers of digital services, especially when they involve cloud infrastructure, health-related data, and vulnerable populations, as is the case with LSP7. It requires that systems implement incident reporting and risk management measures.
- **The Cybersecurity Act (CSA)**[56]: establishes an EU-wide certification framework for ICT products, services, and processes. These certifications can demonstrate compliance with article 15 of the AI Act.
- **The Cyber Resilience Act (CRA)**[57]: mandates continuous monitoring, regular updates, and vulnerability management throughout a product's lifecycle.

**Spanish regulations**

- **Clinical-research law & GCP Transposition**
  - **Ley 14/2007**[58], de 3 de julio, de I**nvestigación Biomédica**: requires an ethics-committee approval (within FAITH, approved by CEI UPV), a clear protocol and data-management plan, and annual safety reports.
- **Data protection law**
  - **Ley Orgánica 3/2018**[59], de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales (**LOPDGDD**): Reinforces GDPR's consent

[53] Charter of Fundamental Rights of the European Union [2012] OJ C.
[54] Directive 2009/136/EC of the European Parliament and of the Council of 25 November 2009 amending Directives 2002/22/EC and 2002/58/EC and Regulation (EC) No 2006/2004 [2009] OJ L337/11 (ePrivacy Directive).
[55] Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union [2022] OJ L333/80 (NIS 2 Directive).
[56] Regulation (EU) 2019/881 on ENISA and on information and communications technology cybersecurity certification [2019] OJ L151/15 (Cybersecurity Act).
[57] Proposal for a Regulation of the European Parliament and of the Council on horizontal cybersecurity requirements for products with digital elements and amending Regulation (EU) 2019/1020 COM/2022/454 final (Cyber Resilience Act).
[58] Ley 14/2007, de 3 de julio, de Investigación biomédica [2007] BOE núm. 159, 4 julio, p. 28826–28848.
[59] Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales [2018] BOE núm. 294, 6 diciembre.

requirements and adds the obligation to publish the DPO's contact and data-rights channels.

- **Real Decreto 1720/2007**[60], de 21 de diciembre, por el que se aprueba el Reglamento de desarrollo de la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal (**RLOPD**): Specifies security measures by "levels", where health data falls under "high level". This implies pseudonimization and encrypted transport requirements.
- **Ley 41/2002**[61], de 14 de noviembre, básica reguladora de la **autonomía del paciente** y de derechos y obligaciones en materia de **información y documentación clínica**: Complements LOPDGDD/GDPR for health data. Stipulates patients' right to information, autonomy in consenting, and clinical-documentation standards (such as large-font plain-language text, especially relevant for elderly people).

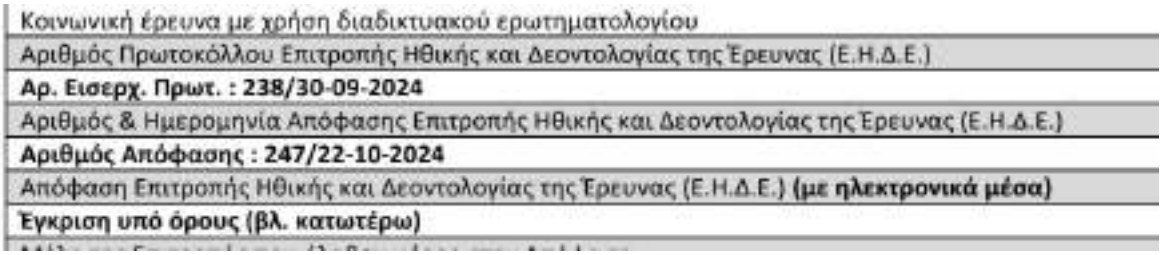
[60] Real Decreto 1720/2007, de 21 de diciembre, por el que se aprueba el Reglamento de desarrollo de la Ley Orgánica 15/1999, de 13 de diciembre, de protección de datos de carácter personal [2008] BOE núm. 17, 19 enero.
[61] Ley 41/2002, de 14 de noviembre, básica reguladora de la autonomía del paciente y de derechos y obligaciones en materia de información y documentación clínica [2002] BOE núm. 274, 15 noviembre.

# 3. Stakeholder engagement and survey analysis

## 3.1. Stakeholder identification and engagement strategy

The FAITH project places great emphasis on human-centered engagement, to ensure that diverse voices are included in shaping the trustworthiness of AI systems. Given the ethical and regulatory complexities surrounding AI, engaging with stakeholders from different domains is not just a necessity—it is a responsibility. FAITH acknowledges that trust in AI cannot be built in isolation; it must emerge through open dialogue, critical reflection, and a shared commitment to responsible innovation. To this end, a structured stakeholder identification and engagement strategy was developed, designed to capture perspectives from across all seven Large-Scale Pilots (LSPs) and incorporate them into the broader framework of AI trustworthiness. A cornerstone of this strategy was the use of anonymous surveys, allowing stakeholders to share their insights, concerns, current practices, and expectations without fear of attribution or bias. Anonymity was essential in fostering honest and uninhibited feedback, particularly on sensitive issues such as AI transparency, fairness, reliability, and compliance with regulatory frameworks like GDPR and the AI Act. The survey targeted several stakeholders. By reaching across disciplines and professional backgrounds, FAITH ensured that the survey results reflected a rich diversity of experiences and perspectives on AI trustworthiness.

Ethical considerations were paramount throughout the stakeholder engagement process. The survey and interview protocols were assessed and approved by the Research Ethics Committee of FORTH (Figure 4 shows the confirmation protocol's number and decision) to ensure compliance with European legal and ethical standards, including GDPR. Participants were fully informed about the purpose of the study, how their data would be used, and their rights regarding participation. No personally identifiable information was collected, reinforcing the commitment to privacy and confidentiality. Moreover, participants were reminded not to disclose any personal or sensitive data in their responses, further safeguarding their anonymity. These measures were not just about legal compliance; they were about fostering a culture of respect, trust, and responsibility—values that underpin FAITH's mission. The Research Ethics Committee of FORTH is organized based on the provisions of Law 4521/2018, and the current regulatory framework. The mission of the Research Ethics Committee is to reassure that the quality of the research projects conducted at the FORTH stand at the highest standards in terms of bioethical issues, according to the National and European Law. Specifically, the scope of the Research Ethics Committee at FORTH is to assess the compliance of a research project with the European Legislation on Bioethics in respect to the value of humans, the autonomy of the participants, to their private life and their personal data, as well as the natural and cultural environment. Funded research programs contacted by members of FORTH that include clinical or laboratory research in humans, experiments with human samples,), research on animals or the environment (natural and cultural), is mandatory to be submitted for approval to the Research Ethics Committee.


Κοινωνική έρευνα με χρήση διαδικτυακού ερωτηματολογίου

Αριθμός Πρωτοκόλλου Επιτροπής Ηθικής και Δεοντολογίας της Έρευνας (Ε.Η.Δ.Ε.)

**Αρ. Εισερχ. Πρωτ. : 238/30-09-2024**

Αριθμός & Ημερομηνία Απόφασης Επιτροπής Ηθικής και Δεοντολογίας της Έρευνας (Ε.Η.Δ.Ε.)

**Αριθμός Απόφασης : 247/22-10-2024**

Απόφαση Επιτροπής Ηθικής και Δεοντολογίας της Έρευνας (Ε.Η.Δ.Ε.) **(με ηλεκτρονικά μέσα)**

**Έγκριση υπό όρους (βλ. κατωτέρω)**

**Figure 4.** An instance from the 1st FAITH anonymous survey's confirmation protocol provided by the Research Ethics Committee at FORTH.

The recruitment of stakeholders was carefully structured and coordinated through action plans developed by each LSP, ensuring comprehensive and representative participation. LSP coordinators played a pivotal role in identifying key stakeholders relevant to their domain and facilitating outreach efforts. Direct invitations were supplemented by multi-channel distribution strategies, including institutional mailing lists, professional networks, and research collaborations, ensuring broad and meaningful engagement. Recognizing that AI trustworthiness is a context-dependent issue, the survey was adapted to reflect sector-specific concerns, making it relevant and accessible to different stakeholder groups.

## 3.2. Objectives of the survey

FAITH's commitment to trustworthy and human-centric AI requires a deep understanding of how different stakeholders perceive, interact with, and evaluate AI systems. To achieve this, FAITH conducted its **1st Anonymous Survey**, an initiative designed to systematically assess the perspectives of AI practitioners, domain experts, policymakers, and end-users across the seven Large-Scale Pilots (LSPs). This survey serves as a foundational step to identify key challenges, expectations, and priorities related to AI trustworthiness to ensure that the activities related to the FAITH_AI_TAF are aligned with real-world needs and concerns. The primary goal of the survey was to gain a deeper insight into the current practices, attitudes, and concerns regarding responsible and trustworthy AI. Specifically, the survey aimed to:

- Explore the strategies employed by AI developers to enhance transparency, fairness, privacy, and security in AI models.
- Evaluate the level of understanding that different stakeholder groups have about AI systems, their functionalities, and associated risks.
- Assess the acceptance of AI across various sectors and identify factors that influence trust in AI-driven decision-making.
- Investigate stakeholders' perspectives on regulatory compliance, ethical considerations, and accountability mechanisms.
- Identify existing gaps and barriers that hinder the development, deployment, and societal adoption of AI technologies.

## 3.3. Survey implementation

To facilitate broad participation and to ensure ease of access, the FAITH survey was implemented using EU Survey[62], an official European Commission tool for conducting secure and GDPR-compliant online surveys. The EU Survey platform was chosen for its: (i) strong data protection measures, to ensure that responses remain anonymous and secure, (ii) user-friendly interface, to make participation seamless for stakeholders across different backgrounds, (iii) scalability, to enable outreach across multiple sectors and geographic regions. The survey was disseminated through multiple channels, including institutional networks, LSP leaders, and targeted outreach efforts to ensure a well-represented sample size. Towards this direction, the survey participants received well-defined instructions along with an overview of the survey's objectives before proceeding, reinforcing transparency and ethical integrity in data collection.

[62] European Commission. (n.d.). EU Survey. Retrieved June 25, 2025, from https://ec.europa.eu/eusurvey/home/welcome.

The survey was carefully structured to balance comprehensiveness with ease of participation, ensuring that respondents could provide meaningful input without undue burden. It was divided into three groups of equations: (i) the *general group of questions*, (ii) the *domain specific group of questions*, (iii) the *group of legal and ethical* questions. Each group includes a combination of multiple-choice questions, Likert-scale assessments, and open-ended questions, allowing for both quantitative analysis and qualitative insights. This hybrid approach was adopted to ensure a richer understanding of AI trustworthiness from both technical and societal perspectives. The purpose of the survey along with the objective, confidentiality policy, communication, contact and additional information regarding standards/policies/legislations are presented in Annex A. This information is included as an introductory section of the FAITH 1st Anonymous survey. Figure 5 illustrates the initial setup interface of the EU Survey tool's "Editor" panel. It shows how the introductory content of the survey was configured to inform participants about the FAITH project goals and the purpose of the stakeholder engagement process.



**Figure 5.** An instance from the EUSurvey ("Editor" panel, Part 1 - Introduction).

Figure 6 displays the survey design interface for entering and structuring the questionnaire items. It captures how specific questions were created to gather input on AI trustworthiness, stakeholder expectations, and perceptions in alignment with FAITH's co-creation objectives.



**Figure 6.** An instance from the EUSurvey ("Editor" panel, Part 2 - Questions).

Figure 7 shows the testing environment within the EU Survey platform used to validate the user experience before survey launch. It highlights how the FAITH team ensured clarity, accessibility, and GDPR-compliant design through internal testing.




**Figure 7.** An instance from the EUSurvey ("Test" panel).

## 3.4. Sample size estimation

In 2024, the number of individuals aged 20 to 64 in the EU was approximately 197,566,000[63], where 0.41% of EU workers are AI professionals (estimated at approximately 810,000 individuals[64]). Based on these findings and considering the exploratory nature of the study, a sample size calculation was performed assuming a 95% confidence level and a ±10% margin of error, following standard survey design guidelines[65,66]. Under these parameters, a minimum number of approximately 96 valid responses was deemed sufficient. The decision to adopt a ±10% margin of error reflects the survey's intended use for mapping general attitudes and strategies rather than for supporting statistical hypothesis testing or precise inferential conclusions. The achieved sample size of 107 therefore satisfies this threshold. As such, the response pool is considered robust for the purpose of exploratory analysis, subgroup comparison, and knowledge co-creation within the scope of the FAITH project. The formula that was used to estimate the sample size (introduced by Cochran[65]) was defined as in:

$$n = \frac{Z^2 p(1-p)}{E^2}, \quad (1)$$

where $Z$ is the z-score corresponding to the desired confidence level (equal to 1.96 for 95% confidence), $p$ is the estimated population proportion (equal to 0.5 for maximum variability), $E$ is the desired error margin (equal to 0.1, expressed in decimal form). Substituting the chosen parameters in (1) yields the following minimum number of samples:

$$n = 1.96^2 \times 0.5 \times (1-0.5) \,/\, 0.1^2 = 0.9604/0.01 = 96.04 \sim 96.$$

## 3.5. Survey analysis workflow

According to Figure 8, the analysis of the FAITH 1st Anonymous Survey followed a structured, multi-step workflow to ensure a rigorous, transparent, and reproducible approach. The process combined

[63] https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Employment_-_annual_statistics.
[64] https://economicgraph.linkedin.com/content/dam/me/economicgraph/en-us/PDF/AI-in-the-EU-Report.pdf.
[65] Cochran, William Gemmell. Sampling techniques. john wiley & sons, 1977.
[66] https://www.qualtrics.com/blog/calculating-sample-size/.

quantitative and qualitative methods, leveraging advanced data science and statistical modeling techniques to extract meaningful insights. By utilizing both Python for exploratory and visualization tasks and R for Bayesian analysis, the workflow aimed to uncover patterns, correlations, and key findings that would inform AI trustworthiness requirements and stakeholder perspectives.

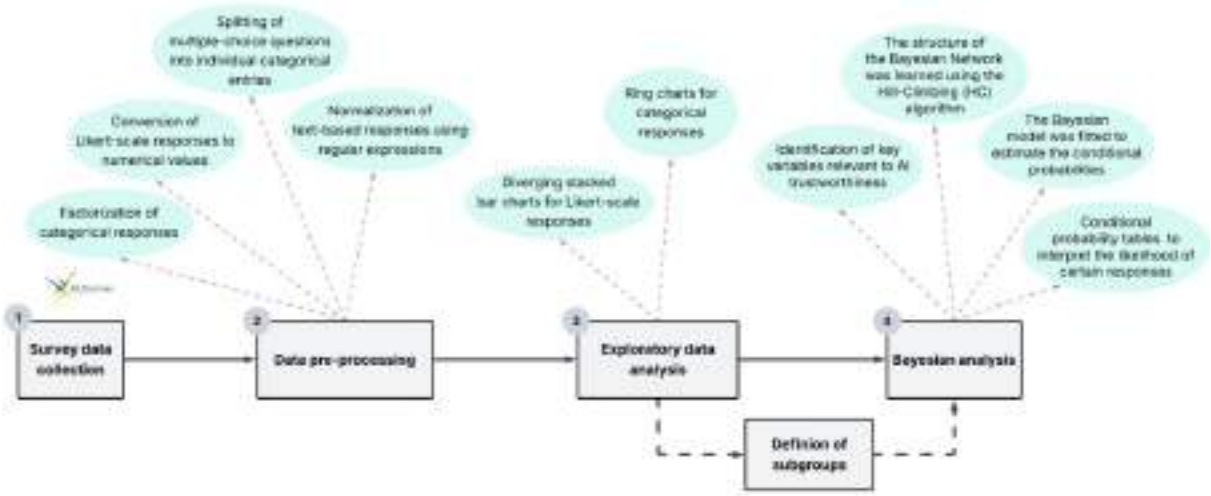



**Figure 8.** The survey analysis workflow.

*3.5.1. Survey data collection*

The survey was hosted and managed on the EU Survey platform. Once the survey period was completed, the raw dataset was exported in a .csv format (43 questions, 107 samples).

*3.5.2. Data pre-processing*

The dataset underwent systematic preprocessing to prepare it for data exploration. Likert-scale responses were converted to numerical values for statistical processing. Categorical responses were factorized to allow subgroup analysis. Multiple-choice questions (where respondents could select multiple options) were split into individual categorical entries for accurate visualization. Text-based responses were normalized using regular expressions to ensure consistency in wording and labeling. Responses were validated to confirm they aligned with expected scales (e.g., 1-5 or 1-7 Likert scales). The dataset did not include any responses with incomplete data in critical fields because the fields were mandatory.

*3.5.3. Exploratory data analysis*

To uncover patterns, stakeholder trends, and key insights, an exploratory analysis was performed using seaborn and matplotlib in Python 3.12. Emphasis was given on data visualization to make the responses more interpretable, particularly for a diverse audience comprising AI experts, policymakers, and domain professionals. Diverging stacked bar charts were created for Likert-scale questions to offer a clear breakdown of stakeholder sentiment—from skepticism to full acceptance of AI. Ring charts (donut plots) were used to visualize categorical responses in order to reflect the distribution of stakeholder backgrounds, roles, and expertise levels within the FAITH project.

*3.5.4. Bayesian analysis*

While descriptive statistics and visualizations provided valuable high-level insights, FAITH also sought to uncover deeper patterns and relationships within the data. To achieve this, Bayesian Network

analysis was conducted using the R "bnlearn" library[67], allowing for a probabilistic assessment of how different factors influence stakeholder perspectives.

**The goal of the Bayesian analysis is to explore the interdependencies between stakeholder perceptions of AI trustworthiness, understanding and acceptance.**

To this end, the Bayesian analysis followed a structured approach as described below:

- Key variables relevant to AI trustworthiness were first identified, including:
  - the most critical trustworthiness dimension that the stakeholders consider (**Q2.1**),
  - the stakeholder understanding of AI (**Q2.6**), and
  - the stakeholder acceptance of AI (**Q2.7**).
- The Tabu Search metaheuristic optimization algorithm[68] was utilized to learn the structure of the Bayesian Network in order to facilitate the detection of probabilistic relationships between different survey responses. The Tabu Search optimization algorithm was selected to overcome the limitations of traditional methods (like the Hill-Climbing[69]) which might get stuck in local optima. Finally, the structure of the Bayesian model was optimized using the Akaike Information Criterion (AIC) scoring method.
- The Bayesian model was fitted to allow for the estimation of conditional probabilities in order to reveal how one response variable influences another.
  - For instance, to determine whether a stakeholder's level of understanding of AI had a direct impact on their trust and acceptance of AI-driven decisions.
- Conditional probabilities were extracted to interpret the likelihood of certain responses occurring based on other factors.

### *3.5.5. Subgroup analysis*

Subgroup analysis was conducted to examine how the responders' perspectives varied between different subgroups. More specifically, two subgroups were defined:

- Subgroup 1. The responders who work in "Research Centres" (75 samples).
- Subgroup 2. The responders who have 11+ years of expertise in the domain (42 samples).

### *3.5.6. Tools and technologies used*

To support the overall analysis, a combination of Python and R programming languages was used to leverage state-of-the-art data science and statistical modeling libraries, as described next.

- **Python** libraries used for exploratory data analysis and visualization:
  - *pandas*[70]*, NumPy*[71]: For data pre-processing and transformation.
  - *seaborn*[72]*, matplotlib*[73]: To visualize the survey responses and to ensure clear and effective communication of insights.
- **R** libraries used for Bayesian analysis:

[67] Scutari, Marco, Maintainer Marco Scutari, and Hiton-PC MMPC. "Package 'bnlearn'." Bayesian network structure learning, parameter learning and inference, R package version 4.1 (2019).
[68] Glover, Fred, and Manuel Laguna. Tabu search. Springer US, 1998.
[69] Hinson, John M., and J. E. R. Staddon. "Matching, maximizing, and hill-climbing." Journal of the experimental analysis of behavior 40.3 (1983): 321-331.
[70] https://pandas.pydata.org/
[71] https://numpy.org/
[72] https://seaborn.pydata.org/
[73] https://matplotlib.org/

- *bnlearn*[74]: To learn the structure of the Bayesian Networks and estimate conditional probabilities.
- *ggplot2*[75]: To visualize probabilistic relationships and interactions between variables.
- *openxlsx*[76]: To export the Bayesian Networks analysis results in a structured format.

## 3.6. Survey analysis results

The current section presents the detailed survey analysis results. In a nutshell, the survey analysis reveals several important insights regarding stakeholder thoughts, priorities, and challenges in the development and deployment of trustworthy AI-based systems. The respondents, primarily from research institutions and with substantial domain expertise, identified transparency, reliability, and ethical compliance (e.g. fairness, data privacy) as core dimensions of trustworthiness. While many participants focus on explainability tools, sensitivity analysis, and data anonymization, challenges remain in bias mitigation, regulatory alignment, and stakeholder engagement (co-design approaches). Importantly, participants rated their own understanding and acceptance of AI systems at approximately 3 out of 5, suggesting a moderate level of familiarity that leaves room for further education, dialogue, and co-design. This underscores the importance of making AI-based systems not only technically sound but also understandable and approachable. There was also strong interest in tools like the FAITH AI model passport, confirming that the FAITH AI_TAF is well aligned with stakeholder needs for transparency, explainability, and responsible AI governance. These findings provide an empirical foundation for further tailoring the FAITH framework to real-world needs across the LSPs.

### *3.6.1. Exploratory data analysis results (Overall)*

*General group of questions*

**Q1.1 Respondent's Profile (select one or more options)**

The majority of the survey respondents (Figure 9) have a background in Computer Science or Information Engineering (34.7%), followed by Data Science (26.5%) and Engineering (22.9%). Smaller proportions represent LSP Domain Experts (8.8%) and Mathematicians (7.1%), reflecting a diverse but technically focused respondent pool.

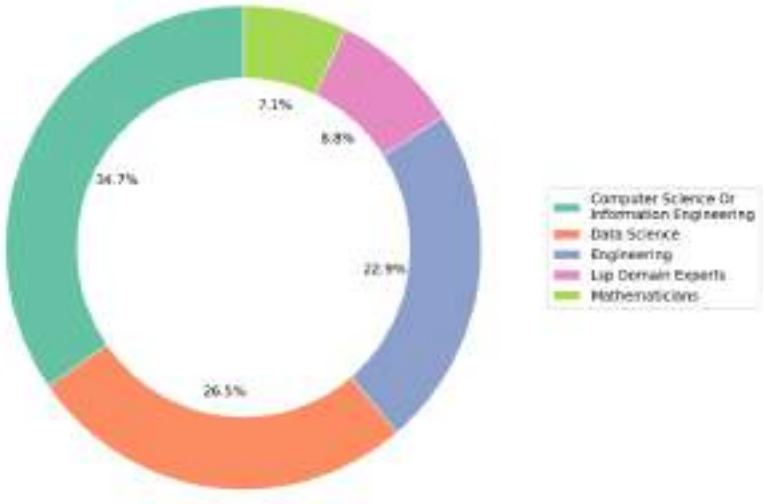



**Figure 9.** Distribution plot of the responses from Q1.1.

[74] https://www.bnlearn.com/
[75] https://ggplot2.tidyverse.org/
[76] https://cran.r-project.org/web/packages/openxlsx/index.html

**Q1.2 Years of expertise in the domain? (select only one option)**

According to Figure 10, the majority of respondents (39.3%) reported having over 11 years of experience in their field, followed by 25.2% with 2–5 years, 22.4% with 5–10 years, and 13.1% with 0–2 years. This indicates that the survey captured insights from a predominantly experienced cohort of stakeholders.

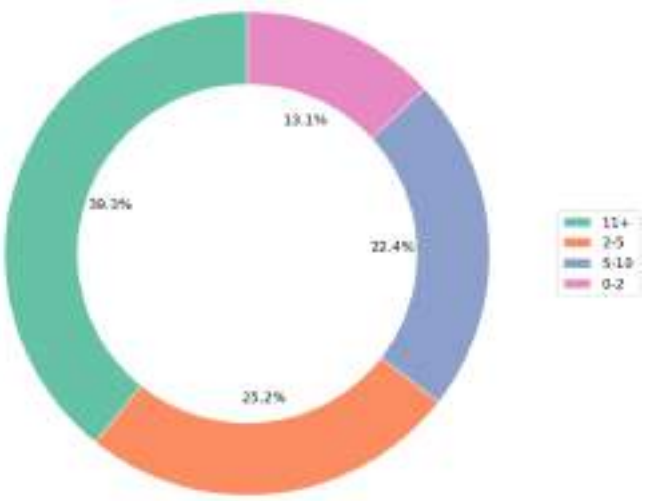



**Figure 10.** Distribution plot of the responses from Q1.2.

**Q1.3 What type of organization do you work in? (select one or more options)**

According to Figure 11, the majority of survey respondents (63.6%) are affiliated with research centres, followed by 18.6% from public organizations. Respondents from private companies in specific sectors accounted for 9.3%, while 8.5% represented tech companies. This distribution highlights the strong involvement of academic and public-sector stakeholders in the FAITH project.

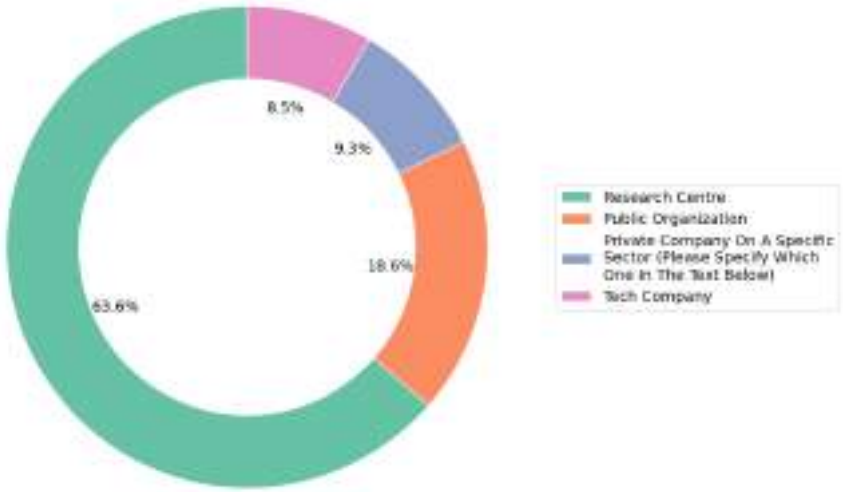



**Figure 11.** Distribution plot of the responses from Q1.3.

**Q1.4 If you selected Private Company in the previous question, please provide the answer here**

The respondents who work in a Private Company represented a balanced mix of stakeholder domains. Education accounted for the largest share at 25%, while each of the remaining domains, including Mass Media, Healthcare, Media, Journalism Saptamana.Online, Pia Petros Ion Lucian (Robotics), and FORTH, contributed equally with 12.5%. This diverse distribution ensures broad coverage across FAITH's pilot domains.

**Q1.5 What approach do you adopt in AI developing process to ensure transparency? (select one or more options)**

The respondents emphasized various practices for improving model transparency (Appendix B, Supplementary Figure 1). The most frequently selected approach was precise documentation of model training (34.0%), followed by MLOps practices (23.9%) and model sharing on GitHub (22.6%). Model cards (10.1%) and other approaches (9.4%) were less commonly cited. These insights reflect a strong preference for traceability and reproducibility in AI development.

**Q1.6 If you selected Other in the previous question, please provide the answer here**

The responses to the other approaches include open-source sharing, model passports, XAI tools (e.g., SHAP), no-code platforms, experiment tracking tools, and stakeholder involvement (Appendix B, Supplementary Figure 2). The even distribution suggests a broad and balanced appreciation across various transparency and traceability practices in AI development.

**Q1.7 What types of reliability metrics of the developed AI models in your domain do you utilize? (select one or more options)**

The most cited key practices for ensuring model robustness and reliability (Appendix B, Supplementary Figure 3) were sensitivity analysis (25.4%) and generalization capacity on external datasets (24.9%), followed by post-deployment performance monitoring (18.0%) and uncertainty analysis and quantification (14.6%). Robustness to adversarial examples (13.2%) and other methods (3.9%) were also mentioned, indicating a broad awareness of critical evaluation techniques across the AI lifecycle.

**Q1.8 If you selected Other in the previous question, please provide the answer here**

The respondents mentioned that they also use other reliability metrics (Appendix B, Supplementary Figure 4) like SHAP/LIME, self-explanatory architectures, feature importance analysis, as well as conducting small-scale research prototypes. The even distribution suggests no single approach dominates current practices across the surveyed group.

**Q1.9 What methods or practices does your organization use to analyze and detect harmful bias in AI-based systems? (select one or more options)**

According to Figure 12, 40.2% of respondents reported conducting pre-deployment bias testing, while 34.4% indicated they have no current practices in place for addressing bias. Only 25.4% reported engaging in post-deployment monitoring. This suggests that while proactive bias mitigation is somewhat prioritized, ongoing monitoring and comprehensive practices remain limited.

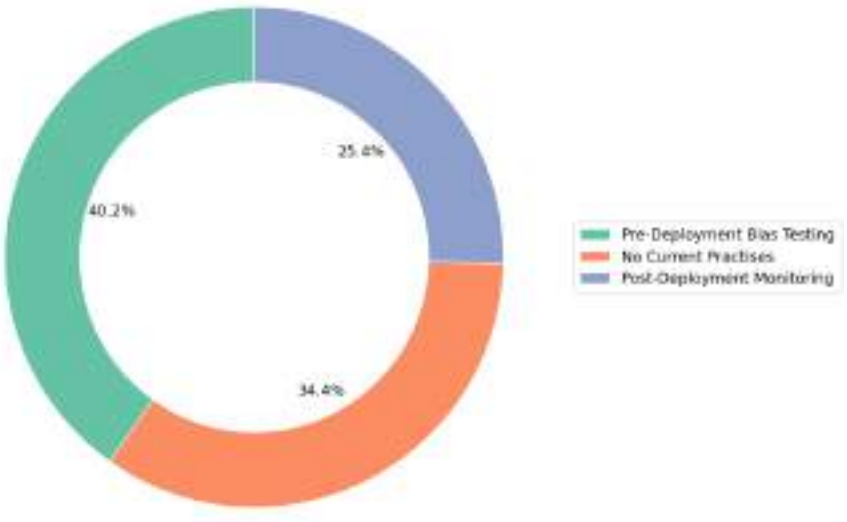



**Figure 12.** Distribution plot of the responses from Q1.9.

**Q1.10 What are the ethical concerns you care the most during the AI lifecycle? (select one or more options)**

Data privacy (16.2%) and data fairness (15.4%) were identified as the top trustworthiness concerns among respondents (Figure 13). These were followed by model fairness (13.2%), system robustness and generalization capacity (13.0%), and system and model transparency (11.9%). Lower but still notable attention was given to data security (10.8%), system safety and system and model security (both 6.8%), and system decision opacity (5.9%). This distribution highlights a predominant focus on ethical and quality-related dimensions of trust.

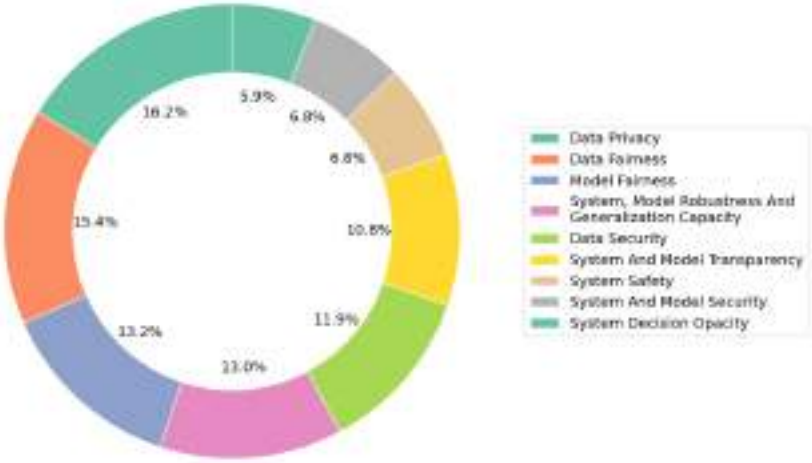



**Figure 13.** Distribution plot of the responses from Q1.10.

**Q1.11 How much time do you spend with the end-users/domain experts to define and test your AI system? (select only one option)**

According to Figure 14, 38.3% of respondents reported engaging with AI systems for between 1 and 6 months, followed by 26.2% for less than one month, 20.6% for 6 to 12 months, and 15.0% for more than one year. This distribution indicates that although a significant number of respondents are in the early stages of user engagement (less than six months) there is a meaningful segment (approximately 35.6%) who engage with domain experts over more extended periods (6 months to over a year). This highlights a diversity in maturity levels of AI development processes across the surveyed participants.

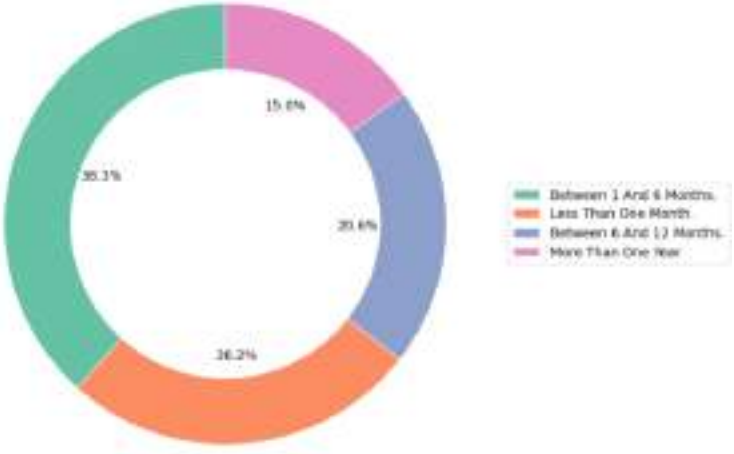



**Figure 14.** Distribution plot of the responses from Q1.11.

**Q1.12 How do you secure the privacy threat of your input dataset? (select one or more options)**

Data anonymization techniques were the most commonly cited approach (47.8%) for privacy-preserving AI development (Appendix B, Supplementary Figure 5). This was followed by offline learning (24.5%), synthetic data generation techniques (17.6%), and federated learning approaches

(10.1%). The responses indicate a strong emphasis on data protection strategies, particularly through **anonymization.**

**Q1.13 Can you provide the above mentioned methods as open-source code? (select only one option)**
50.5% of respondents reported that they cannot provide the mentioned methods as open-source code, while 49.5% stated they can (Appendix B, Supplementary Figure 6). This nearly even split highlights a significant limitation in the open-source availability of methods used within the community.

**Q1.14 How does your organization assess and communicate the risks associated with AI-systems to end-users and other interested stakeholders (e.g. domain experts, business leaders, security teams, quality assurance engineers)? (select one or more options)**
The most preferred approach for increasing stakeholder trust in AI systems (Appendix B, Supplementary Figure 7) is through detailed documentation and reporting (38.9%). This is followed by holding regular briefings and training sessions (22.3%), transparency tools like AI model cards or explainability features (21.1%), and risk assessment frameworks (12.6%). Only 5.1% selected other methods. These results highlight a strong emphasis on clear communication and structured documentation to foster trust.

**Q1.15 If you selected Other in the previous question, please provide the answer here**
The responses regarding the use of other methods to assess and communicate the risks associated with AI-systems to end-users and other interested stakeholders, were evenly distributed (Appendix B, Supplementary Figure 8). This indicates a highly diverse range of practices regarding how risks related to AI systems are communicated and managed—ranging from formal risk assessments and ethical oversight to informal acknowledgment of risks or project-level handling during prototyping.

**Q1.16 How often does your organization engage in co-creation activities with end-users and other interested stakeholders to align AI system development with their needs and concerns? (select one or more options)**
According to Figure 15, the most common practice for addressing trustworthiness in AI systems is at key stages of the development process (32.4%). This is followed by regular and consistent efforts throughout the project lifecycle (27.3%) and occasional integration based on specific project needs (23.7%). Fewer respondents reported rarely or never addressing trustworthiness (10.8%) or considered it not applicable (5.8%). These results indicate a general trend toward structured and integrated trustworthiness efforts across development phases.

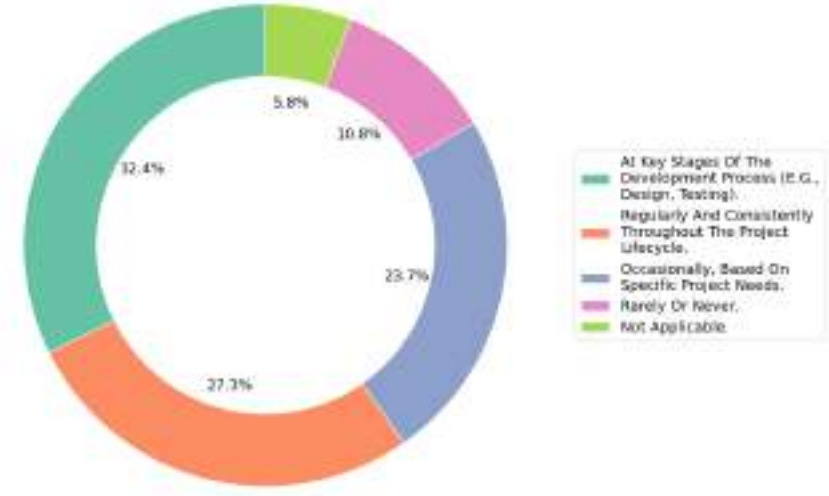



**Figure 15.** Distribution plot of the responses from Q1.16.

**Q1.17 How does compliance with regulatory frameworks (e.g., GDPR, AI Act) influence your organization's AI system design and deployment strategies? (select only one option)**

According to Figure 16, 37.4% of respondents indicated that compliance with regulatory frameworks like GDPR and AI-Act influences some parts of the process, particularly data handling, while 32.5% stated it has a significant influence, guiding all aspects of design and deployment. A smaller portion (11.4%) consider it mainly during final deployment stages, and only 7.3% report it has a minimal influence. Just 5.7% selected not applicable or mainly an afterthought, suggesting that most organizations acknowledge regulatory compliance as a substantial design factor, though the level of integration varies.

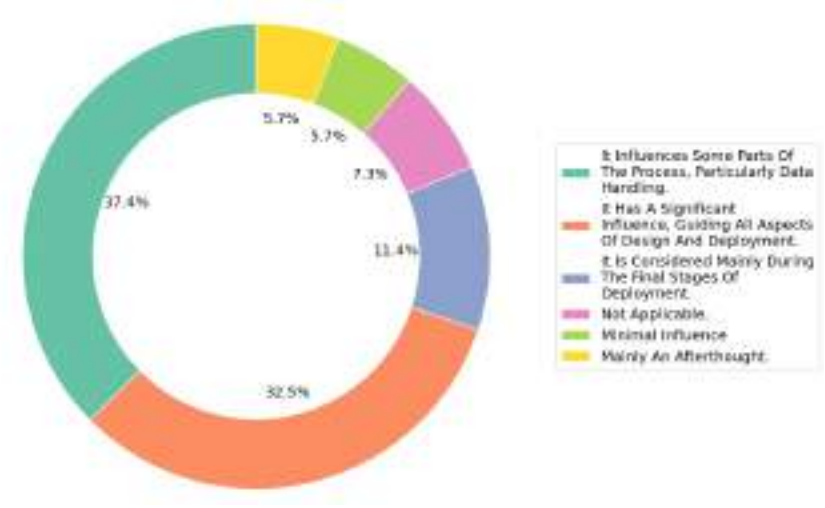



**Figure 16.** Distribution plot of the responses from Q1.17.

**Q1.18 What do you see as the most significant challenges in achieving trustworthy AI in your sector? (select one or more options)**

The top challenge identified by respondents is enhancing transparency and explainability of AI models (25.8%) (Figure 17). This is followed closely by ensuring data privacy and security (23.2%) and managing and mitigating biases in AI systems (23.2%). Other notable challenges include keeping up with regulatory changes and compliance requirements (15.5%) and aligning AI systems with ethical and societal values (12.4%). These results highlight a multidimensional concern where technical, ethical, and legal challenges are all prominent in the pursuit of trustworthy AI.

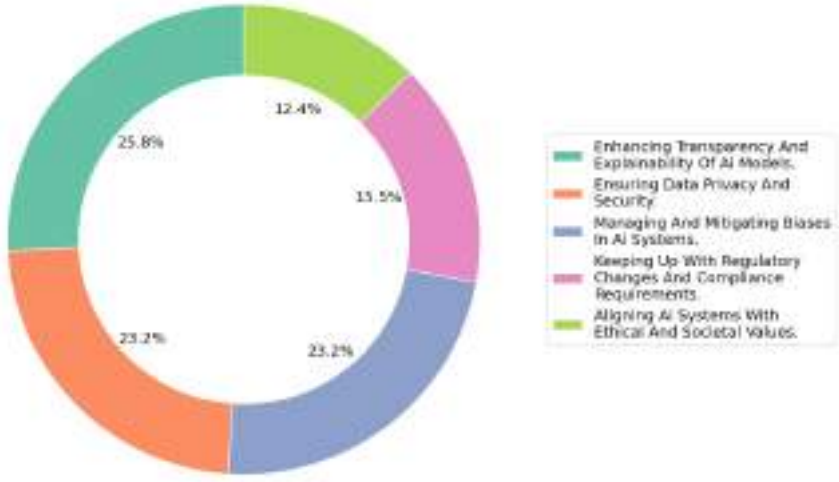



**Figure 17.** Distribution plot of the responses from Q1.18.

**Q1.19 How does your organization measure success in addressing critical Trustworthy dimensions such as fairness, transparency, and robustness in your AI systems? (select one or more options)**

According to Figure 18, the most common method for measuring success in trustworthy AI is through predefined KPIs and performance metrics (39.2%). This is followed by user feedback and satisfaction surveys (26.8%) and continuous monitoring and evaluation post-deployment (22.9%). A smaller share of respondents (11.1%) indicated reliance on third-party audits and certifications. These results suggest a strong emphasis on internal measurement frameworks, complemented by user perspectives and ongoing system evaluation.

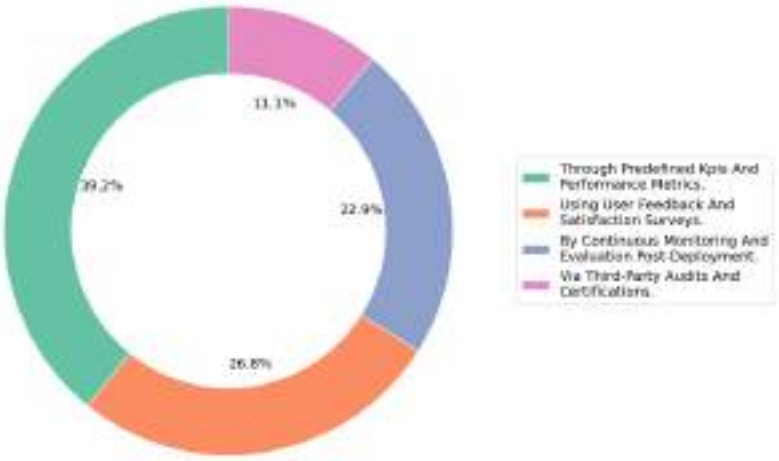



**Figure 18.** Distribution plot of the responses from Q1.19.

**Q1.20 Approximately how much time does your team spend collaborating with domain experts during the AI model design and testing phases? (select only one option)**

In total, 32.7% of participants spend approximately 1–3 months collaborating with domain experts during the AI model design and testing phases (Appendix B, Supplementary Figure 9). Other durations reported include 3–6 months (19.6%), 6–12 months (17.8%), more than one year (16.8%), and less than one month (13.1%).

**Q1.21 How does your organization ensure continuous improvement and adaptation of AI models to changing data and environments? (select one or more options)**

According to Figure 19, the most common approach is implementing a continuous feedback loop with users and stakeholders (30.1%), followed by regularly retraining models with new data (27.9%). Other strategies include collaborative R&D with external experts (19.1%), conducting periodic audits and assessments (13.1%), and using automated monitoring systems to detect and address performance drifts (9.8%). These results reflect a balanced mix of technical, collaborative, and user-driven strategies for model adaptability.

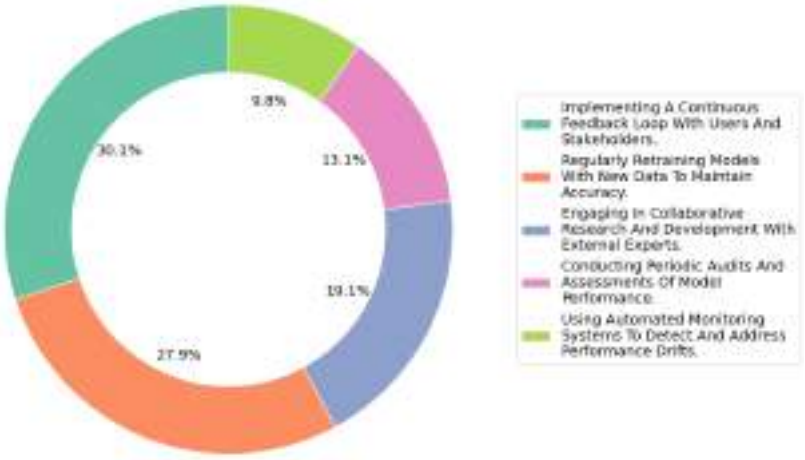



**Figure 19.** Distribution plot of the responses from Q1.21.

**Q1.22 What ethical guidelines and governance frameworks does your organization follow to ensure responsible AI development and deployment? (select one or more options)**

According to Figure 20, the most frequently cited approach is following the European Commission's Ethics Guidelines for Trustworthy AI (33.5%). This is followed by implementing internal ethics review boards or committees (19.8%), and both adhering to international standards (e.g., IEEE) and engaging with external ethicists for independent reviews, each at 15.6%. Smaller proportions reported no current ethical governance practice (8.4%) or conducting regular ethics training (7.2%). These results suggest that while EU-level guidelines are prominent, there remains variability in formalized internal or external ethical oversight.

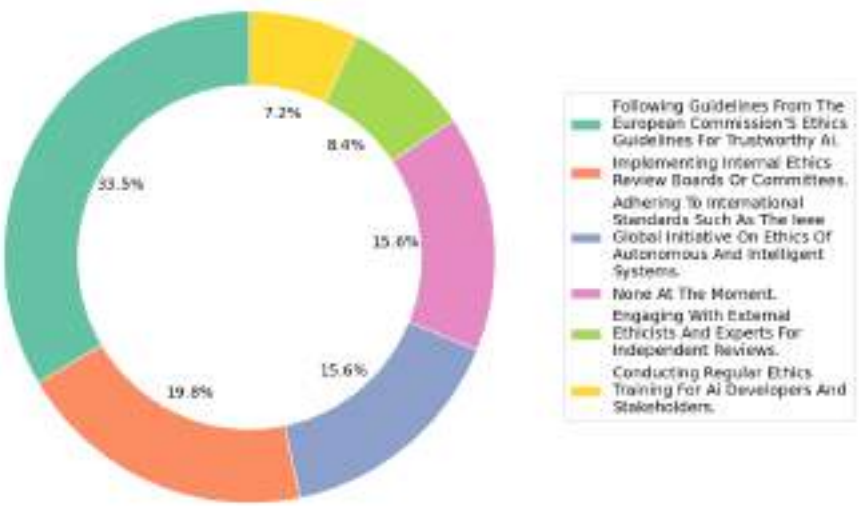



**Figure 20.** Distribution plot of the responses from Q1.22.

**Q1.23 Do you believe that the FAITH AI model passport can be used for the monitoring of the AI systems in your organization? (select only one option)**

According to Figure 21, 64.5% of respondents believe that the FAITH AI model passport can be used for monitoring AI systems in their organization. Meanwhile, 30.8% indicated uncertainty ("I don't know"), and only 4.7% responded "No". These results indicate a generally positive perception of the FAITH model passport, although a substantial portion of participants remain unsure about its applicability.

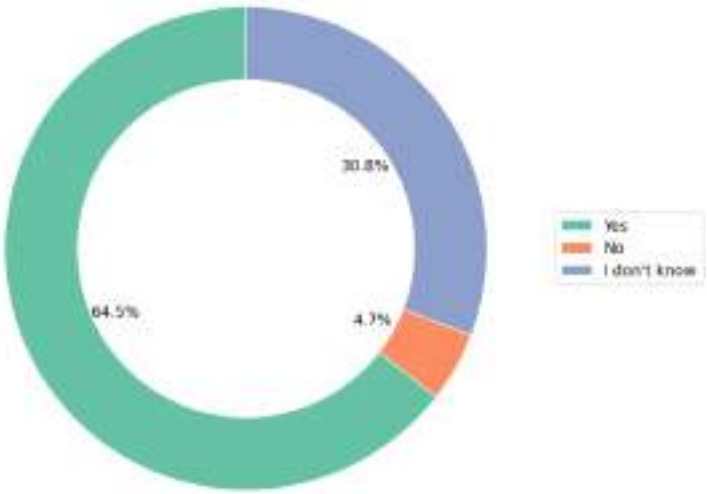



**Figure 21.** Distribution plot of the responses from Q1.23.

*Note: The FAITH AI model passport monitors, tracks and lists, in a standardised format (i.e. as metadata), all the model's details (e.g. models' scope, actors and developers involved, development*

*tools and technical choices made, versions, performance metrics, and various trustworthiness characteristics), as well as, all the details of the data used to train it (e.g. data provenance and localization).*

**Q1.24 Do you believe that SpyderRisk can be useful for assessing AI based systems in your organization? (select only one option)**

In total, 50.5% of respondents believe that SpyderRisk can be useful for assessing AI-based systems in their organization (Appendix B, Supplementary Figure 10). 44.9% were uncertain, responding "I don't know", and only 4.7% answered "No". This indicates a generally favorable but cautious attitude toward the tool, with a notable portion of stakeholders yet to form a definitive opinion.

*Note: Spyderisk is for assessing risk in complex socio-technical systems. By complex systems we mean "systems of systems that humans are unable to fully understand, debug or predict, typically socio-technical-physical systems." In other words, humans build critical systems that are unknowable, and so we need to apply mathematics to delineate the risks in these unknowable systems. Link: https://github.com/Spyderisk.*

**Q1.25 How does your organization ensure that AI models are understandable to end-users and stakeholders, including providing explanations for decisions made by the AI? (select one or more options)**

According to Figure 22, the use of ad-hoc explainability methods applied to the AI models (32.2%) is the most common approach to ensure that AI models are understandable to end-users and stakeholders, followed by the selection of explainable-by-design methods (28.3%). Describing processes directly to stakeholders without specific automated methods was chosen by 23.7% of respondents, while 15.8% indicated that no explainability methods are currently in use. These results reflect a growing but uneven adoption of strategies to ensure AI explainability for end-users.

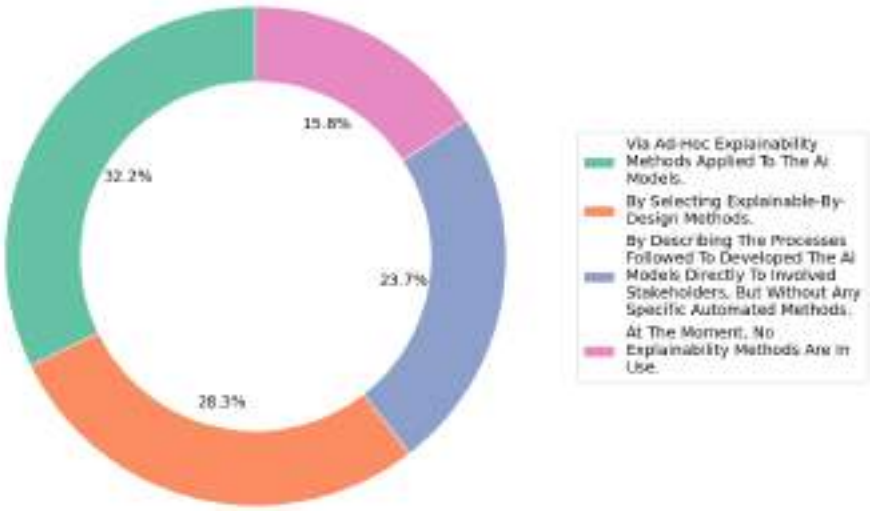



**Figure 22.** Distribution plot of the responses from Q1.25.

**Q1.26 What measures does your organization take to assess and mitigate biases in AI systems to ensure fair treatment?**

Respondents indicated a range of practices adopted by their organizations. Several participants reported that their organizations conduct thorough analyses of training data to identify and remove potential biases, with some also undertaking dedicated audits of datasets to detect imbalances. Balancing datasets by collecting additional data for underrepresented groups was highlighted as another common strategy. In some cases, organizations integrate causality into AI models to address underlying bias-related factors more robustly. The generation of synthetic or targeted data to reduce bias was also mentioned as a mitigation approach. Furthermore, some respondents noted that

experimentation related to bias and fairness is evaluated within the framework of the Data Protection Impact Assessment (DPIA). However, a subset of participants acknowledged that their organizations do not currently implement specific procedures for bias assessment or mitigation in the development or deployment of AI systems.

**Q1.27 How does your organization test and validate AI systems to ensure they can handle a variety of inputs and operate reliably under different conditions? (select one or more options)**
The most common method reported is sensitivity analysis (43.3%), followed by sandbox testing (27.0%) and stability analysis (22.0%) (Appendix B, Supplementary Figure 11). A smaller proportion of respondents (7.8%) selected other methods. These findings indicate that most organizations prioritize proactive testing strategies to evaluate system behavior under varying conditions.

**Q1.28 If you selected Other in the previous question, please provide the answer here**
The other methods included the use of external datasets, multiple external datasets, performance evolution checks, test data validation, user testing, and performance metrics (Appendix B, Supplementary Figure 12). Some respondents also indicated that no formal standards or certifications are currently applied. This even distribution suggests a diverse and non-standardized approach to real-world AI performance evaluation across organizations.

**Q1.29 How does your organization safeguard AI systems against malicious attacks and ensure the integrity and confidentiality of the data they process? (select one or more options)**
The most commonly reported security measure is restricted internet access (54.9%), followed by adversarial training and testing (42.7%). Only 2.4% of respondents indicated that no measures are currently in place (Appendix B, Supplementary Figure 13). These results reflect a strong awareness of AI system security, with most organizations employing proactive defenses against potential threats.

**Q1.30 What processes are in place within your organization to monitor AI system performance, address any issues that arise, and ensure accountability for the outcomes produced by AI? (select one or more options)**
In total, 62.9% of respondents reported using in-house methods to monitor AI systems in operation, while 36.1% rely on tools such as Grafana, Prometheus, ML Flow, and similar solutions (Appendix B, Supplementary Figure 14). Only 1.0% indicated that they do not currently use any monitoring methods. These results suggest that most organizations prioritize custom or tailored monitoring solutions to ensure effective oversight of their AI systems.

**Q1.31 How does your organization ensure that AI systems consistently perform as intended and deliver accurate and dependable results over time? (select one or more options)**
According to Supplementary Figure 15 (Appendix B), 55.6% of participants reported using sandbox testing to ensure that their AI systems consistently perform as intended and deliver accurate and dependable results over time. Meanwhile, 44.4% utilize formal verification methods. These results suggest that while both approaches are employed, sandbox testing is slightly more prevalent among respondents.

**Q1.32 Does your organization provide to the AI engineers guidelines and good practices for responsible AI development? (select only one option)**
According to Figure 23, 42.1% of participants stated that their organization provides AI engineers with guidelines and good practices for responsible AI development. However, a combined 57.9% either do not (33.6%) or don't know (24.3%) if such support is offered, indicating that more than half of organizations may lack formalized or communicated frameworks for responsible AI.

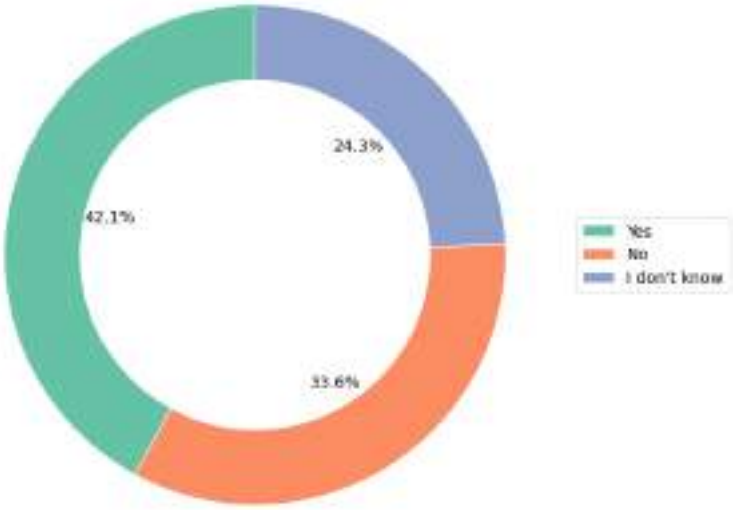



**Figure 23.** Distribution plot of the responses from Q1.32.

*Domain specific group of questions*

**Q2.1 What is the most critical Trustworthiness dimension from your working organization/pilot/domain? (select one or more options)**

The most critical dimension of trustworthiness identified by participants is Explainability and Interpretability (21.1%), followed by Fairness (18.7%) and Robustness and Reliability (14.4%) (Figure 24). Other notable dimensions include Privacy Preservation (12.4%), Safety (12.4%), Transparency and Accountability (11.7%), and Security (9.4%). These results reflect a broad prioritization across technical, ethical, and societal aspects of trustworthy AI.

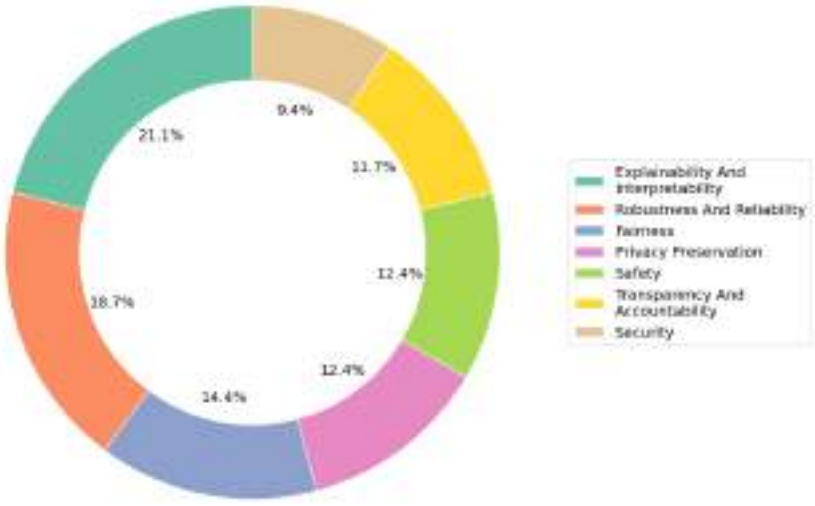



**Figure 24.** Distribution plot of the responses from Q2.1.

**Q2.2 Have you implemented anything to address these critical dimensions, if it is quantifiable? (select only one option)**

According to Figure 25, 69.2% of the participants reported that they haven't implemented anything to address the identified critical trustworthiness dimensions. In contrast, 30.8% stated that they have implemented measures. This suggests a significant implementation gap, with the majority of organizations yet to act on these priorities despite recognizing their importance.

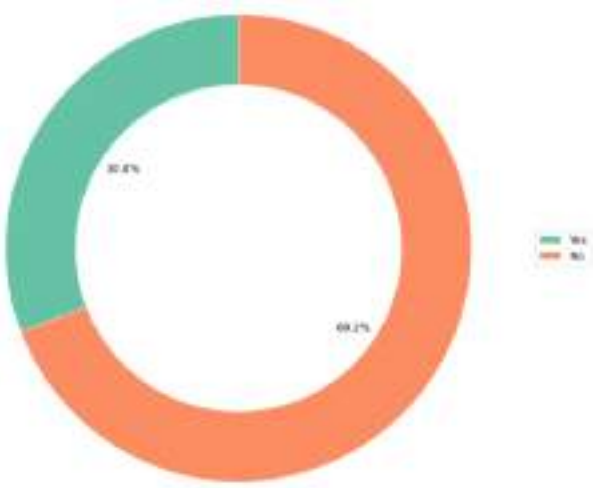


**Figure 25.** Distribution plot of the responses from Q2.2.

**Q2.3 If you answered yes to the previous question, please specify the method you have implemented**

Several participants described the use of in-house development, including custom-built explainability tools, MLOps frameworks, and bias assessment mechanisms tailored to their specific systems. Others rely on open-source tools and frameworks, benefiting from community-driven resources for transparency and reproducibility. Adversarial robustness testing was also cited, helping ensure systems remain stable under malicious input. Privacy-preserving techniques such as anonymization, de-identification of personal data, and metrics reporting were noted for maintaining data confidentiality. Some organizations apply causality integration and explainable-by-design architectures to enhance system interpretability and accountability from the outset. Additionally, practices like fair performance testing and bias analysis were employed to evaluate fairness in model outputs. A subset of respondents reported aligning with FUTURE-AI principles, a framework for promoting ethical and responsible AI research and development. Lastly, the use of generative AI and Large Language Models (LLMs) was mentioned as part of the technical landscape, indicating the integration of cutting-edge AI capabilities within their trustworthy AI initiatives.

**Q2.4 Is your method open-access? (select only one option)**

In total, 63.6% of respondents stated that their method is not open access, while only 36.4% confirmed that it is (Appendix B, Supplementary Figure 16). This indicates that the majority of participants operate with proprietary or restricted-access approaches, which may pose challenges for transparency, reproducibility, and wider adoption of trustworthy AI practices.

**Q2.5 What are the relevant legislation, standards, policies (for the AI based systems) in your organisation?**

Participants cited a mix of regulatory frameworks and institutional guidelines. The General Data Protection Regulation (GDPR) was the most frequently mentioned, reflecting its foundational role in data protection across the EU. Several respondents also identified the EU Artificial Intelligence Act (AI Act) as a key legislative framework shaping AI development and deployment. For organizations operating in the healthcare domain, the Medical Device Regulation (MDR) in conjunction with GDPR was highlighted as particularly relevant. Some participants referred to internal or national policies, which provide organization-specific or country-specific rules for responsible AI use. International standards, notably ISO/IEC 27001 (information security management) and ISO/IEC 24028 (AI trustworthiness), were also acknowledged as part of formal compliance strategies. Outside the EU context, the Health Insurance Portability and Accountability Act (HIPAA) was mentioned, indicating adherence to U.S.-based health data protection requirements. However, a number of respondents

(nine) admitted they did not know what policies applied to AI within their organization, signaling a gap in awareness or training. Additionally, seven respondents reported that no formal policies are currently in place, pointing to the absence of structured governance for AI systems in some settings.

**Q2.6 Are your involved stakeholders in AI-based systems understand AI? (select from 1-5 scale, 1 being not at all)**

The majority of the respondents (45.8%) rated stakeholder understanding of AI at level 3, indicating a moderate level of understanding (Figure 26). This was followed by level 2 (18.7%), level 4 (16.8%), level 5 (10.3%), and level 1 (8.4%). These results suggest that while most stakeholders possess a basic to moderate grasp of AI concepts, there remains room for enhanced education and engagement.



**Figure 26.** Distribution plot of the responses from Q2.6.

**Q2.7 Are your involved stakeholders in AI-based systems accept the AI? (select from 1-5 scale, 1 being not at all)**

According to Figure 27, 33.6% of participants indicated that stakeholders in their AI-based systems accept AI at a level 3 out of 5, suggesting moderate acceptance. This is followed by level 2 (27.1%), level 4 (19.6%), level 5 (13.1%), and level 1 (6.5%). These findings indicate that while full acceptance is limited, most stakeholders are at least somewhat receptive to AI integration.



**Figure 27.** Distribution plot of the responses from Q2.7.

**Q2.8 Identify the key considerations for trustworthy & human acceptable AI based systems in your organization**

Participants emphasized a range of technical, ethical, and regulatory priorities. Foremost among these was the need to ensure that AI systems are reliable and free from harmful biases, highlighting fairness as a fundamental requirement. Respondents also stressed the importance of building systems that are both understandable and secure, underscoring the dual necessity of transparency and protection

against threats. Another critical consideration was the traceability of AI decisions, with an emphasis on ensuring that system operations are safe and auditable. Making AI systems comprehensible to users and stakeholders was highlighted as essential for fostering trust and acceptance. Compliance with established regulations, particularly GDPR, was noted as a key standard for data privacy and performance accountability. Finally, participants pointed to the importance of developing AI systems that perform consistently across diverse conditions, reflecting a focus on robustness and generalizability as part of trustworthy AI design.

*Group of legal and ethical questions*

**Q3.1 How often are legal and ethics experts involved in the development or deployment of AI systems? (select only one option)**

Based on Supplementary Figure 17 (Appendix B), 22.4% of participants indicated that legal and ethics experts are frequently involved, followed closely by occasionally involved (21.5%) and when necessary (18.7%). Others reported that experts are rarely involved (16.8%), never involved (11.2%), or always involved (9.3%). These results show that while expert involvement is present in many cases, consistent integration throughout the AI lifecycle is still lacking.

**Q3.2 What is your opinion on the involvement of legal and ethics experts in the AI development and deployment process? (select only one option)**

The majority of respondents (34.6%) view the involvement of legal and ethics experts as essential for ensuring compliance and accountability (Supplementary Figure 18, Appendix B). Others believe it is important but could be more streamlined (29.9%), or useful but not always necessary (23.4%). Smaller percentages see involvement as minimal and should be increased (6.5%), not required (5.6%), or that it is not currently involved (not directly indicated but inferred through low percentages). These insights highlight a strong recognition of the value of expert involvement, with opportunities to improve consistency and integration.

**Q3.3 What legal concerns and risks do you consider most important throughout the AI lifecycle? (select one or more options)**

According to Figure 28, the most important legal concerns identified by participants are compliance with data protection and privacy laws (e.g., GDPR) (18.2%) and fairness and non-discrimination in AI models (16.0%). Other key concerns include violation of fundamental rights (13.3%), transparency and accountability for decision-making processes (12.8%), and compliance with sector-specific regulations (11.7%). Lesser-cited risks include cybersecurity and model security (10.6%), liability risks and system safety/failures (10.3%), and compliance with anti-discrimination laws (7.1%). These findings highlight a multidimensional legal risk landscape with a primary focus on data privacy and fairness.

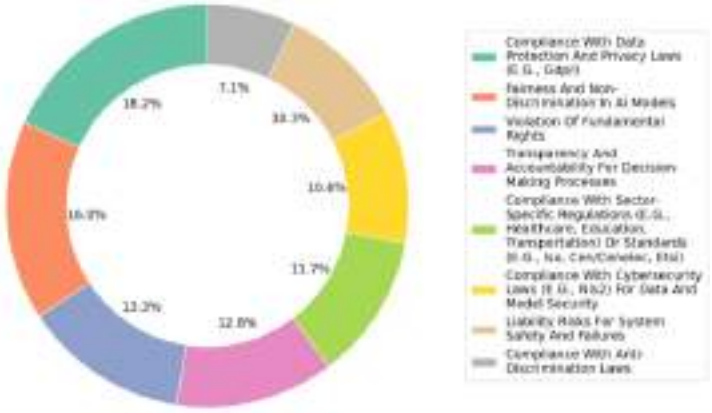



**Figure 28.** Distribution plot of the responses from Q3.3.

*3.6.2. Exploratory data analysis results on Subgroups 1 and 2*

This section presents the exploratory analysis results for the two subgroups defined in Section 3.5.5:

- Subgroup 1. The responders who work in "Research Centres" (75 samples).
- Subgroup 2. The responders who have 11+ years of expertise in the domain (42 samples).

*3.6.2.1. Key findings from Subgroup 1*

Participants from research centres showed a relatively higher emphasis on adopting structured documentation practices for AI development (Appendix B2, Supplementary Figures 19-23). Over one-third reported using precise documentation of model training and version control mechanisms such as MLOps pipelines or GitHub repositories for model transparency. Additionally, this group frequently cited the use of advanced model explanation tools like SHAP (SHapley Additive exPlanations), model cards, and AI model passports. In terms of trustworthiness considerations, respondents from research centres often ranked explainability and interpretability, followed by fairness, as the most critical dimensions. Despite this awareness, a significant portion of this group indicated that their organizations had not yet implemented formal mechanisms to operationalize these principles, citing resource and integration challenges. Finally, Subgroup 1 respondents appeared more engaged in monitoring practices, reporting greater use of post-deployment performance tracking and sandbox testing environments to evaluate system reliability and accountability over time.

*3.6.2.2. Key findings from Subgroup 2*

Subgroup 2 displayed an in-depth understanding of AI lifecycle risks and often brought attention to cross-cutting legal and ethical concerns (Appendix B3, Supplementary Figures 24-28). This group was more likely to reference a combination of standards, including GDPR, the AI Act, and ISO/IEC standards. Their responses indicated a stronger orientation toward compliance-by-design and proactive risk assessment (e.g., use of bias analysis tools and adversarial robustness testing). Experienced professionals also demonstrated more critical engagement with stakeholder dynamics, emphasizing the importance of stakeholder understanding and acceptance of AI systems. A majority of respondents in this subgroup rated stakeholder understanding as moderate (3 out of 5), aligning with the broader sample, yet they were more likely to advocate for structured stakeholder education programs and legal-ethical integration within AI development teams. Interestingly, Subgroup 2 participants expressed greater skepticism toward AI systems' social readiness, often indicating that explainability efforts were insufficient or poorly communicated, which could hinder trust even in technically robust systems.

*3.6.3. Bayesian analysis results (Overall)*

According to Figure 29, Q2.1 (critical trustworthiness dimensions) is an independent factor, since it exhibits varied preferences across the respondents. Stakeholder **acceptance** (Q2.7) is **highly influenced** by their **understanding** (Q2.6) of AI-based systems. Through our findings we identify the need to improve **stakeholder understanding** (Q2.6) to **enhance acceptance** (Q2.7) of AI-based systems.




**Figure 29.** The structure of the estimated Bayesian model.

*3.6.4. Bayesian analysis results on Subgroups 1 and 2*

*3.6.4.1. Key findings from Subgroup 1*

Bayesian posterior distributions suggest that members of Subgroup 1 were significantly more likely to:

- Value explainability and transparency as top trustworthiness priorities,
- Use advanced documentation practices, such as model versioning and structured metadata,
- Adopt model explainability tools (e.g., SHAP, LIME, or model cards),
- Favor open science practices, such as publishing model performance and decision criteria.

Posterior probabilities indicated with >85% confidence that this subgroup was more inclined to prioritize traceability and openness compared to other populations. However, the analysis also revealed a lower likelihood (around 60%) of formal integration of legal frameworks or ethical risk assessments into technical workflows. This may reflect a research-driven environment focused more on innovation and methodological development than on operational compliance.

*3.6.4.2. Key findings from Subgroup 2*

Subgroup 2 exhibited a markedly different profile. Bayesian estimates showed with high probability (>90%) that this group was:

- More likely to incorporate legal and ethical frameworks, particularly GDPR, ISO/IEC standards, and preliminary AI Act provisions,
- More focused on risk mitigation and stakeholder impact, with increased attention to fairness, accountability, and downstream effects,
- Less likely to rely on purely technical mitigation tools, instead favoring hybrid strategies involving human oversight and documentation controls.

In contrast to Subgroup 1, these respondents demonstrated greater confidence in evaluating stakeholder needs, with posterior probability distributions showing a 75–85% likelihood of endorsing participatory or co-design approaches as essential for trust.

Interestingly, Bayesian comparison also indicated a lower-than-average probability (<60%) that these stakeholders used model explainability tools in practice. This suggests a potential disconnect between high-level ethical awareness and the routine use of technical methods for interpretability, possibly due to contextual or institutional constraints.

## 3.7. Respondents' understanding of AI

The majority of respondents rated their understanding of AI as moderate. On a 5-point Likert scale, the median response was 3, suggesting that while most participants are aware of AI's general capabilities and risks, fewer feel confident in comprehensively understanding the inner workings, data requirements, or design decisions behind AI tools. This result was consistent across professional sectors, although slightly higher self-assessed understanding was observed among respondents affiliated with research institutions and those with more than 11 years of domain experience (as also confirmed in Sections 3.6.2 and 3.6.4). In open-ended responses, several participants expressed concern over limited transparency and explainability of current AI systems, which they described as a barrier to developing deeper understanding. A recurring theme was the need for contextualized information, such as domain-specific case studies or practical demonstrations, to help stakeholders interpret AI decisions and limitations more effectively. Respondents also emphasized the importance of training and continuous learning opportunities tailored to their role, especially in high-impact domains such as healthcare, education, and transport. Another key insight was the discrepancy

between perceived AI competence and actual exposure to AI tools. Some respondents indicated high familiarity with AI terminology but limited practical engagement with deployed systems. This highlights the importance of distinguishing between conceptual awareness and applied understanding when assessing readiness for AI adoption. The findings suggest that stakeholder understanding is an essential, but currently uneven, foundation for trustworthy AI deployment. FAITH's engagement strategy should continue to prioritize knowledge transfer, hands-on training, and the integration of explainability mechanisms tailored to end-user needs. These insights directly inform the co-design of the FAITH AI_TAF, particularly in the development of human-centric indicators of trustworthiness and the integration of knowledge accessibility as a key trust-enhancing measure.

## 3.8. Respondents' acceptance of AI

The average response to Q2.6 across participants was approximately 3 out of 5, indicating that stakeholders are generally perceived to have a moderate understanding of AI systems. However, this level of understanding may not be sufficient to support full and confident acceptance of AI in practice. Several participants highlighted a disconnect between technical development and stakeholder comprehension, particularly in non-technical or user-facing domains. This gap in understanding was frequently associated with hesitation or conditional acceptance—especially when AI system outputs were perceived as opaque or misaligned with the specific values and expectations of a given sector. Further support for this relationship between understanding and acceptance is found in responses to Q1.18, where a significant portion of participants identified transparency and explainability as key challenges to achieving trustworthy AI. Stakeholders are seen as more willing to engage with AI systems when they clearly understand their function, limitations, and implications for human decision-making. In contrast, a lack of explainability, concerns over diminished human autonomy, and ambiguous accountability structures were consistently cited as factors that limit broader acceptance. While 58% of respondents rated stakeholder understanding of AI at a moderate level (3 out of 5) in Q2.6, qualitative and subgroup analysis suggests domain-specific differences. Stakeholders in research and technology-oriented environments—such as those represented in Subgroup 1—were perceived as having greater exposure to AI tools and methodologies. These participants reported higher adoption of documentation and transparency practices, and expressed more openness to AI integration, particularly where oversight and relevance to user needs were clearly communicated.

## 3.9. Respondents' support preferences and interaction with AI

This section explores the preferences, expectations, and concerns of stakeholders regarding their interaction with AI systems, as identified through the FAITH anonymous survey and subsequent analysis. Stakeholders—especially those with technical backgrounds—express a clear demand for systems that are not only legally compliant but also intelligible, responsive, and human-aligned. The analysis highlights explainability as a foundational requirement, not just for transparency but as a core component of usability and trust-building. In parallel, stakeholders emphasized the need for AI to augment rather than replace human decision-making, underscoring the importance of adjustable autonomy and human oversight. Support mechanisms should be personalized, context-aware, and accessible to users of varying digital literacy levels. Stakeholders also noted that trust is deeply contextual—shaped by application domain, perceived safeguards, and the degree to which AI behavior aligns with human expectations. Bayesian insights from the survey (Section 3.6.3) confirm a strong dependency between understanding and acceptance, reinforcing that trustworthy AI must be designed from the user's cognitive and emotional standpoint. This section advocates for a shift toward explainable-by-design, human-centric systems, and presents the key dimensions of interaction preferences (Table 9).

**Explainability as a Core Expectation (from Q1.18):** Stakeholders explicitly demand explainability—not just for compliance, but as a practical interface for trust-building. This aligns with the *AI Act* (Article 13) and standards like *ISO/IEC 24028*, which emphasize human understanding of system behavior. Explainability isn't just a post-hoc need; it's foundational to interaction design. From a technical standpoint, this supports the use of: AI Model Passport, Google Model cards and AI FactSheets; SHAP, LIME, GRADCAM++, or counterfactuals for local explanations; and dashboards for real-time traceability**.** Such preference also implicitly supports the integration of **Explainable-by-Design (XbD)** models over ad-hoc interpretability methods.

**Support vs. Autonomy Balance:** There is a need to calibrate autonomy: stakeholders want AI that *augments—not reg.places—human decision-making.* This directly relates to *Human Oversight* requirements in the AI Act (Annex III risk categories) and *Trustworthy AI* pillars defined by HLEG. Thus, systems must implement **interactive interfaces** with adjustable autonomy levels and the decision outputs they provide should always be auditable and reversible. A logging human-AI interactions (e.g., override frequency) can act as a safety metric.

**Personalization of Support:** Stakeholders show interest in **adaptive** support that aligns with individual user needs. Technically, this means that such systems should integrate **user Profiling** into AI UX layers. Secondly exploit **Federated Learning** or other privacy techniques for tailoring responses without compromising data privacy. Lastly, AI systems should provide **multi-modal explanations** (textual, visual, or voice-based) depending on user capabilities. Overall End Users/Stakeholders would like to push for *context-aware trustworthiness mechanisms*, linking to robustness, inclusiveness, and accessibility principles.

**Human-Centered Expectations: Stakeholders expect AI systems** don't act alone but co-exist with human actors. Stakeholders expect these systems to be *accountable, d*emonstrate *situational awareness* (not static logic), and offer **empathy** or socio-emotional cues, especially in healthcare or education.

**Anxiety and Ambiguity Around Autonomy:** From an SSH lens, support preferences reflect a deeper cultural tension: stakeholders are cautious of **machine agency**. This is reminiscent of **technological paternalism**, where humans are reluctant to cede control—even if AI might make "better" decisions.

**Digital Inequalities and Usability:** A profound ethical implication is that interaction preferences are shaped by **digital literacy**, **age**, **accessibility**, and **domain knowledge**. AI systems must consider these *intersectional variables* when designing support structures: e.g., explanations for a medical AI should differ for a clinician vs. a caregiver.

Exploiting the analysis conducted for the FAITH anonymous survey we can provide insights into what are the stakeholders (with a technical background) *qualitative preferences* regarding how they want to **interact** with AI systems. Bayesian analysis in Section 3.6.3 reveals a direct dependency: **stakeholder acceptance (Q2.7)** is heavily influenced by their **understanding of AI systems (Q2.6)**. This quantitatively supports the point from the current section that stakeholders are not ready to engage meaningfully unless systems are designed with **explainability and intelligibility** at their core. The latter justifies from a technical design standpoint, that such AI systems shall integrate: Interactive tutorials; adaptive explanation modules (tailored to stakeholder expertise); uncertainty communication mechanisms

**Explainability Is Valued but Under-Delivered:** Subgroup 1 (Research Centre participants) and Subgroup 2 (Senior experts) reported widespread **adoption of explainability tools** (e.g. SHAP), but also **recognized that these were insufficient for building trust**. This gap mirrors the challenge identified here: explainability tools exist, but **how they're communicated and operationalized** is often misaligned with stakeholder cognitive models. Trustworthy AI frameworks must treat **"Explainability as Communication"**—not just visualization or technical transparency.

**AI Lifecycle Governance Is Fragmented:** Stakeholders, especially senior ones, called for better **risk-based governance and compliance-by-design**, but noted that formal mechanisms (e.g., bias testing, impact assessments) are still underutilized or inconsistently applied.

**Trust Is Contextual and Conditional:** Section 3.8 shows that **trust in AI is not uniform**; it depends on: the **application domain** (e.g. high-stakes health vs. low-risk admin); perceived **safeguards** like oversight, fairness, and data protection. Thus human-AI interaction must be **tailored**—a "one-size-fits-all" model of AI support won't work in practice. Trust must be built **from the user's point of view**, not imposed by system-level compliance.

**Explainability ≠ Acceptability:** A critical SSH finding: even when technical explanations exist, **they don't necessarily lead to social acceptance**—because: Explanations may be poorly contextualized; sociocultural biases (e.g., fear of automation) may persist. This calls for **ethics-by-design**: an approach *where support features are co-designed with affected communities, considering their values and fears—not just their information needs.*

**Table 9.** Dimensions of stakeholder preferences in human-AI interaction.

| Dimension | Qualitative Preference | Quantitative Evidence | Integrated Takeaway |
|---|---|---|---|
| **Explainability** | Stakeholders demand clear, accessible AI explanations | Most critical trustworthiness factor; but insufficiently delivered | Move from "explainable AI" to comprehensible-by- design AI" |
| **Autonomy vs. Support** | Users want augmented—not autonomous—AI | Concerns about overreach of AI in sensitive domains | Design interaction layers with **adjustable autonomy levels** |
| **Understanding as Enabler** | Trust grows with transparency | For instance, a Bayesian model predicts acceptance based on conditional probabilities | Embed educational interaction pathways |
| **Diversity of Preferences** | Preferences are context-sensitive | Subgroup analysis confirms domain and experience shape expectations | AI interfaces must adapt to **stakeholder profiles**, not vice versa |

# 4. Co-design workshops to synthesize considerations into AI trustworthiness requirements and metrics

## 4.1. Connection with T2.3

FAITH's co-design workshops adopt a structured, participatory approach towards the identification of trustworthy and human-acceptable AI characteristics across the full AI lifecycle. It is closely aligned with the objectives of Task 2.3, which aims to "identify key considerations & requirements for trustworthy & human acceptable AI over its whole lifecycle". This alignment is reflected in FAITH's strategic decision to involve stakeholders (domain specific experts, data analysts, AI modelers, scientists/developers) from all seven Large-Scale Pilots (LSPs) through individual workshops that explored domain-specific concerns, perceived threats, vulnerabilities, and control measures. These workshops were conducted as the first engagement point to surface stakeholder expectations and contextual challenges in the application of AI systems (Table 10). Their participatory and exploratory nature laid the groundwork for a broader engagement strategy, which includes forthcoming anonymous surveys and ethnographic approaches. The outcomes from these workshops are being synthesized into actionable trustworthiness requirements and metrics, feeding directly into the FAITH Knowledge Base (KB) and the broader Trustworthiness Assessment Framework (AI_TAF). This iterative, bottom-up methodology ensures that the technical and regulatory responses within FAITH are rooted in real-world pilot experiences and stakeholder realities, thereby enhancing both the legitimacy and the utility of the trustworthiness metrics being developed.

**Table 10.** Activities connected with T2.3.

| |
|---|
| ***"Identify key considerations & requirements for trustworthy & human acceptable AI over its whole lifecycle"*** |
| We organized individual workshops with the LSPs to engage them in shaping the FAITH Knowledge Base (KB) focusing on threats, vulnerabilities, control measures, among others. |
| ***"Work with the pilot stakeholders in the first instance to identify what they perceive to be their main challenges"*** |
| Pilot-specific consequences and vulnerabilities related to various trustworthy AI characteristics and AI lifecycle phases were gathered through the workshop discussions. |
| ***"Engagement via a combination of techniques such as anonymous crowdsourced survey, focus group discussion, co-design and ethnographic study"*** |
| The workshops adopted co-design and discussion elements to foster participatory input, and complemented upcoming survey-based strategies. |
| ***"Synthesize considerations into AI trustworthiness requirements & metrics, plus best practice"*** |
| The insights from the workshops are utilized to update, populate and enhance the content and context of the FAITH KB. |

## 4.2. The co-design journey

The core objective of the co-design journey was to iteratively enhance the FAITH Knowledge Base (KB), focusing on identifying threats, vulnerabilities, and control measures associated with AI-based systems. According to Figure 30, between April 11th and May 5th, 2025, a series of structured, pilot-specific workshops were conducted with stakeholders from all seven Large-Scale Pilots (LSPs). These sessions provided a participatory space for domain experts and end-users to express perceived challenges, ethical concerns, and socio-technical risks associated with the development and deployment of AI in their respective domains. Through these engagements, the project was able to capture pilot-specific consequences and vulnerabilities in relation to key trustworthiness characteristics such as privacy, fairness, accountability, and transparency across different AI lifecycle phases. Each LSP workshop served as a checkpoint, with successful engagements completed for LSP6 (Apr. 11), LSP3 (Apr. 15), LSP4 and LSP7 (Apr. 24), LSP2 (Apr. 28), LSP5 (Apr. 30), and LSP1 (May 5). The participatory design approach adopted throughout these sessions ensured that insights were deeply contextual and grounded in lived operational realities. These outcomes are currently being synthesized and will culminate in the final assessment and integration into the FAITH Knowledge Base by May 30th, 2025. This milestone will feed directly into the continuous refinement of the FAITH Trustworthiness Assessment Framework (FAITH AI_TAF), ensuring that its trustworthiness metrics, requirements, and best practices reflect the needs and values of the stakeholders it is intended to serve.

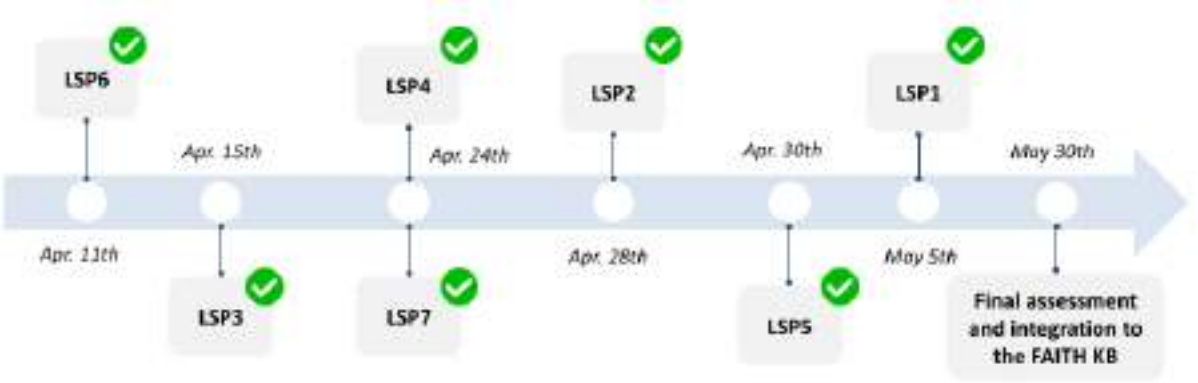



**Figure 30.** The co-design journey.

## 4.3. The phases of the FAITH KB enhancement process

According to Figure 31, the FAITH KB enhancement process was designed around a continuous and iterative cycle that integrates stakeholder engagement, analysis, and synthesis of trustworthiness-related insights into the KB. The process consists of the following 5 phases:

- **Phase 1 - Co-design workshops planning:** This phase involves the strategic coordination of engagement activities with stakeholders from each Large-Scale Pilot (LSP). The planning stage defines objectives, timelines, participant profiles, and thematic focus areas, ensuring that each workshop is tailored to surface relevant risks and trustworthiness concerns.
- **Phase 2 - Co-design workshops execution:** During this phase, the planned workshops are conducted. LSP stakeholders, including AI developers, end-users, and domain experts, participate in structured discussions to share their perspectives on AI trustworthiness challenges, ethical dilemmas, and operational risks.
- **Phase 3 - Identification of vulnerabilities, control measures, threats, and consequences:** Insights gathered during the bilateral workshops are analyzed to identify domain-specific

threats and vulnerabilities, potential control measures, and socio-technical consequences. This phase ensures that the outputs are actionable and aligned with the FAITH trustworthiness dimensions.

- **Phase 4 - Final assessment:** Once all the bilateral workshops with the LSPs have been completed and the KB enhancement templates have been collected (including also the consequences of the related AI-based systems per LSP domain), a thorough review and validation is conducted to identify unique vulnerabilities, control measures and threat categories.
- **Phase 5 - Integration to the KB (Enhancement):** The final, validated outcomes from Phase 4 are formally integrated into the FAITH KB under new IDs. This enhancement step ensures that the KB provides robust, high-quality guidance on threats, controls, and trustworthiness requirements, ready to support AI development and assessment across all pilots.

The FAITH KB enhancement process is a cyclical process that **continuously evolves** across the lifecycle of the project. It guarantees that the FAITH KB is continuously updated to remain aligned with LSPs advancement in order to support the development of trustworthy and human-centric AI systems.

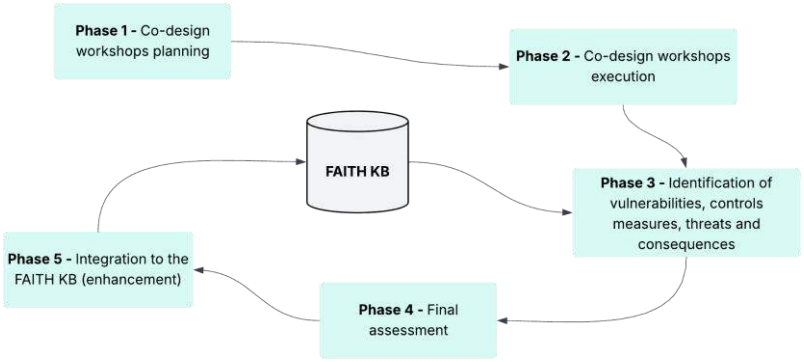



**Figure 31.** The phases of the FAITH KB enhancement process.

## 4.4. Supporting material

Figure 32 presents key definitions included in the FAITH KB enhancement template (Sheet 1), offering clear distinctions between the core concepts of *threat*, *vulnerability*, *control measure*, and *risk*. These definitions provide a shared vocabulary to support stakeholders—both technical and non-technical—in accurately describing AI trustworthiness challenges across pilots. Each term is accompanied by practical examples and an explanation of how they interrelate. For instance, a threat (such as a cyberattack) may exploit a vulnerability (e.g., outdated software), potentially leading to a risk (e.g., data loss), while control measures (like multi-factor authentication) are introduced to reduce or eliminate those risks. By establishing this structured terminology early in the co-design and assessment processes, the FAITH KB ensures a consistent, rigorous approach to analyzing AI-related harms and designing appropriate mitigation strategies.

| Question | Definition | Example | How do they connect? |
|---|---|---|---|
| What is a threat? | A threat is **anything** that has the potential to **cause harm to a system, organization**, or **individual**. It can be **intentional** (e.g., cyberattacks, fraud) or **unintentional** (e.g., natural disasters, human errors). | A hacker attempting to steal data is a threat to an organization's cybersecurity. | A **threat** (e.g., cyberattack) can **exploit a vulnerability** (e.g., weak passwords), **leading** to a **risk** (e.g., data breach). **Control measures** (e.g., enforcing strong passwords) help reduce the risk by addressing vulnerabilities and mitigating threats. |
| What is a vulnerability? | A vulnerability is a **weakness or flaw** in a system, process, or organization that can be exploited by a threat to cause harm. | An outdated software system with security flaws is a vulnerability that hackers can exploit. | |
| What is a control measure? | A control measure (or safeguard) is an **action, process, or technology** put in place to **reduce or eliminate** vulnerabilities, **mitigate** threats, and **minimize** potential damage. | Implementing multi-factor authentication (MFA) is a control measure to protect against unauthorized access. | |
| What is a risk? | A risk is the **potential** for **loss, damage**, or **harm** that could occur when a threat exploits a vulnerability, considering the likelihood and impact of the event. | If a company's database has weak encryption (vulnerability) and cybercriminals target it (threat), the potential loss of sensitive data represents a risk to the company. | |

**Figure 32.** Useful definitions provided in the FAITH KB enhancement template (Sheet 1).

Figure 33 presents a structured matrix from the FAITH KB enhancement template (Sheet 2), illustrating how trustworthiness characteristics—based on NIST definitions—are contextualized through real-world consequences and corresponding control actions. The table uses the healthcare domain (LSP6) as a reference scenario, demonstrating the specific impact of issues such as lack of safety, bias, or privacy failures in AI-based segmentation models. For each of the seven trustworthiness characteristics (e.g., Safe, Secure and Resilient, Fair with bias managed), the table outlines: (1) the formal NIST-based definition, (2) concrete consequences of failing to meet that criterion within the healthcare setting, and (3) recommended control actions or safeguards to address them. For example, under Fair with harmful bias managed, the table notes that biased AI segmentation models may misclassify patients based on demographic factors, leading to unequal diagnostic outcomes. The corresponding control action emphasizes the need for fairness-aware training, bias audits, and representative datasets. This structured mapping ensures that the FAITH KB not only reflects theoretical definitions but also translates them into domain-specific risks and practical mitigation strategies. It supports both contextual relevance and operational guidance, helping pilots and developers apply trustworthiness principles in practice across critical sectors.



**Figure 33.** Useful consequences provided as an example based on the healthcare LSP (LSP6) in the FAITH KB enhancement template (Sheet 2).

Figure 34 shows the structured core form used in Sheet 3 of the FAITH KB enhancement template. This form is designed to guide LSP participants, technical partners, and analysts in systematically capturing and documenting AI-related threats, aligned with the trustworthiness framework established by FAITH. Each row in the form represents a single identified threat, with fields that ensure a complete risk analysis chain from identification to mitigation.



**Figure 34.** The core form for the KB enhancement (Sheet 3).

The form includes the following columns:

- **Threat title**: A concise and descriptive name for the threat.
- **Threat description**: A brief narrative outlining the nature and potential impact of the threat.
- **Affected AI lifecycle stage**: One or more stages in the AI lifecycle (e.g., Design) where the vulnerability is most relevant (Figure 35).
- **Affected AI Asset**: One or more AI assets (e.g., Data) that are vulnerable (Figure 35).
- **Affected TWA dimension/characteristic**: One or more trustworthiness characteristics that are impacted by the vulnerability (e.g., Fair-with harmful bias managed) (Figure 35).

- **Associated vulnerabilities**: Specific weaknesses or flaws that may be exploited by the threat.
- **Associated control measures**: Detailed recommendations for mitigating the threat, including technical or organizational actions.



**Figure 35.** Drop-down options for the Affected AI lifecycle stage, Affected AI Asset, and Affected TWA dimension/characteristic.

## 4.5. Outcomes

In this section, the outcomes of the KB enhancement process, as a result of the broader bilateral workshop meetings, are presented for each LSP.

### *4.5.1. LSP1 – Media*

LSP1 explores the trustworthiness of an AI-based media verification tool designed to support local journalists in evaluating the veracity of emerging narratives before their publication via the PRESShub news portal. The system leverages large language models (LLMs) to analyze and interpret textual information, which introduces specific trustworthiness risks. Key challenges identified include mitigating prompt-based attacks, preventing misinformation propagation, and addressing data provenance and bias issues. Table 11 outlines the seven unique vulnerabilities and associated control measures identified for LSP1, each linked to a corresponding threat, affected AI asset, and trustworthiness (TWA) dimension/characteristic (the completed KB enhancement template can be found in Supplementary Figure 29, Annex C2).

**Table 11.** The identified unique vulnerabilities, related control measures, corresponding threats, affected AI assets, and trustworthiness (TWA) dimension/characteristics for LSP1.

| ID | Associated Vulnerabilities | Associated control measures | Threat title | Affected AI Asset | Affected AI lifecycle stage | Affected TWA dimension/characteristic |
|---|---|---|---|---|---|---|
| LSP1-01 | Prompt Injection | 1. Fine-tune models to ignore override attempts (e.g., "Ignore previous instructions"). 2. Separate user input from system instructions using clear delimiters or metadata. | Prompt attacks | Models | Testing, Production | Valid & Reliable |
| LSP1-02 | Prompt Leakage | 1. Change system prompts per session to | Prompt attacks | Models | Testing, Production | Valid & Reliable |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | reduce predictability. 2. Detect patterns of probing or prompt-revealing questions. | | | | |
| LSP1-03 | Indirect Prompt Injection | 1. Tag or label data sources to inform the model which parts are safe or unsafe to trust. 2. Restrict what autonomous agents can do after reading third-party data | Prompt attacks | Models | Testing, Production | Valid & Reliable |
| LSP1-04 | Hallucination | 1. Apply retrieval-augmented generation (RAG). 2. Estimate token-level probability. 3. Introduce entropy-based measures. | Propagation of misinformation | Models | Design, Testing, Production | Explainable & Interpretable, Valid & reliable |
| LSP1-05 | Data Poisoning | 1. Use curated & vetted training datasets. 2. Recording and trace the origin and lifecycle of the data used to train, fine-tune, or condition the LLM. | Propagation of misinformation | Models | Design, Testing, Production | Explainable & Interpretable, Valid & reliable |
| LSP1-06 | Discrimination and Bias | 1. Bias audits & fairness testing | Regulatory and legal non-compliance | Models, Data | Production | Fair-with harmful bias managed |
| LSP1-07 | Improper Data Usage / Consent Violations | 1. Track data provenance. 2. Engage legal counsel or privacy experts to review: data sources and terms of service. | Regulatory and legal non-compliance | Data | Design, Testing | Fair-with harmful bias managed |

The unique vulnerabilities are distributed across the AI-related lifecycle stages as follows:

- Design – 3 vulnerabilities (LSP1-04, LSP1-05, LSP1-07): Related to reducing hallucinations and safeguarding against data poisoning during model architecture and training.
- Testing – 6 vulnerabilities (LSP1-01, LSP1-02, LSP1-03, LSP1-04, LSP1-05, LSP1-07): Highlighting the importance of evaluating prompt-based attack resistance, data misuse risks, and misinformation robustness.
- Production – 6 vulnerabilities (LSP1-01, LSP1-02, LSP1-03, LSP1-04, LSP1-05, LSP1-06): Emphasizing threats from real-time misuse, bias amplification, and propagation of incorrect outputs in deployed systems.

The unique vulnerabilities are distributed across the AI-related assets as follows:

- Models – 6 vulnerabilities (LSP1-01, LSP1-02, LSP1-03, LSP1-04, LSP1-05, LSP1-06): Involving risks from prompt manipulation, hallucination, bias, and poisoned training data.

- Data – 2 vulnerabilities (LSP1-06, LSP1-07): Addressing improper data usage, lack of consent, and discriminatory content.

The unique vulnerabilities are distributed across the seven TWA dimensions/characteristics as follows:

- Valid & Reliable – 5 vulnerabilities (LSP1-01, LSP1-02, LSP1-03, LSP1-04, LSP1-05): Focused on protecting model integrity against adversarial prompting and misinformation.
- Explainable & Interpretable – 2 vulnerabilities (LSP1-04, LSP1-05): Aimed at enhancing understanding of model behavior and data lineage.
- Fair – With Harmful Bias Managed – 2 vulnerabilities (LSP1-06, LSP1-07): Ensuring legal compliance and minimizing discrimination risks in both data processing and model output.

*4.5.2. LSP2 - Transportation*

LSP2 investigates the trustworthiness of a privacy-preserving AI-based video analytics system aimed at improving safety and efficiency on trains and in stations. The system processes real-time visual data while minimizing personal data storage. Challenges identified include maintaining interpretability, enhancing robustness under real-world conditions, and protecting privacy in a highly sensitive data environment. Table 12 summarizes the three unique vulnerabilities and related control measures as documented within LSP2, each mapped to its corresponding threat, AI asset, and trustworthiness (TWA) dimension/characteristic (the completed KB enhancement template can be found in Supplementary Figure 30, Annex C2).

**Table 12.** The identified unique vulnerabilities, related control measures, corresponding threats, affected AI assets, and trustworthiness (TWA) dimension/characteristics for LSP2.

| ID | Associated Vulnerabilities | Associated control measures | Threat title | Affected AI Asset | Affected AI lifecycle stage | Affected TWA dimension/characteristic |
|---|---|---|---|---|---|---|
| LSP2-01 | Reduced post-hoc interpretability due lack of coupled data-predictions | 1. Provide outputs that include information to support the AI actor responsible for model maintenance, enabling specific model testing even without the coupled input-predictions. These can include details such as the camera that triggered the alert, the location of the object within the camera's field of view (saliency map), and the predicted object label. | Reduced model interpretability | Models | Design, Testing | Explainable & Interpretable |
| LSP2-02 | Identification of specific individuals in the video stream. Collection of data | 1. Image data acquired are not stored, only processed, and only alerts (following | Personal video data acquisition | Data | Testing, Production | Privacy-enhanced |

| | | | | | | |
|---|---|---|---|---|---|---|
| | beyond the intended scope of the model | outputs) along with metadata are sent to the control station. | | | | |
| LSP2-03 | Low robustness to real-world video scenarios | 1. Develop a synthetic dataset—either generative or mixed reality—to incorporate more realistic data for training and testing. 2. Define a DEBUG mode that, in appropriate contexts, allows local video storage. 3. Plan a subsequent testing phase (Phase II) with the train in motion but without passengers (e.g., using the route to the train decoy). | Training data quality due acquisition procedure | Data, Models | Testing | Valid & Reliable |

The unique vulnerabilities are distributed across the AI-related lifecycle stages as follows:

- Design – 1 vulnerability (LSP2-01): Focused on improving model transparency during initial system development.
- Testing – 3 vulnerabilities (LSP2-01, LSP2-02, LSP2-03): Emphasizing the need for robust testing under realistic conditions and strict control of personal data.
- Production – 1 vulnerability (LSP2-02): Addressing privacy risks from real-time deployment and video stream analysis.

The unique vulnerabilities are distributed across the AI-related assets as follows:

- Models – 2 vulnerabilities (LSP2-01, LSP2-03): Related to explainability of model outputs and robustness to real-world video inputs.
- Data – 2 vulnerabilities (LSP2-02, LSP2-03): Concerning protection of personal video data and realistic data acquisition.

The unique vulnerabilities are distributed across the seven TWA dimensions/characteristics as follows:

- Explainable & Interpretable – 1 vulnerability (LSP2-01): Focused on enhancing post-hoc model understanding for maintainability.
- Privacy-enhanced – 1 vulnerability (LSP2-02): Ensuring that personal data are not improperly collected or retained.
- Valid & Reliable – 1 vulnerability (LSP2-03): Highlighting the need for robust performance through improved dataset quality and testing protocols.

#### *4.5.3. LSP3 - Education*

LSP3 assesses the trustworthiness of an AI-based system designed to support student learning through personalized recommendations and performance assessments during laboratory inquiry learning. This

context involves high-stakes interactions between educators, students, and AI systems, making human oversight, fairness, and model reliability essential. Table 13 summarizes the three unique vulnerabilities and related control measures as documented within LSP3, each mapped to its corresponding threat, AI asset, and trustworthiness (TWA) dimension/characteristic (the completed KB enhancement template can be found in Supplementary Figure 31, Annex C3).

**Table 13.** The identified unique vulnerabilities, related control measures, corresponding threats, affected AI assets, and trustworthiness (TWA) dimension/characteristics for LSP3.

| ID | Associated Vulnerabilities | Associated control measures | Threat title | Affected AI Asset | Affected AI lifecycle stage | Affected TWA dimension/characteristic |
|---|---|---|---|---|---|---|
| LSP3-01 | Lack of teachers' training | 1. Communicate to the teachers that their involvement is essential, maintain human oversight. 2. Teachers should be trained how to use the system | Human error | Processes | Production | Safe |
| LSP3-02 | Lack of diverse representation | 1. Implement bias detection and mitigation techniques, regular fairness audits | Bias in Personalized Learning Recommendations | Models | Production | Fair - with harmful bias managed |
| LSP3-03 | Lack of continuous learning | 1. Continuously monitor model performance | Model Drift Leading to Inaccurate Assessments | Models | Production | Valid & Reliable |

The unique vulnerabilities are distributed across the AI-related lifecycle stages as follows:

- Production – 3 vulnerabilities (LSP3-01, LSP3-02, LSP3-03): These reflect post-deployment concerns, including teacher integration, fairness of recommendations, and long-term model performance under real-world use.

The unique vulnerabilities are distributed across the AI-related assets as follows:

- Models – 2 vulnerabilities (LSP3-02, LSP3-03): Highlighting challenges in ensuring fairness in outputs and maintaining model reliability over time.
- Processes – 1 vulnerability (LSP3-01): Emphasizing the need for human oversight and adequate training in operational deployment.

The unique vulnerabilities are distributed across the seven TWA dimensions/characteristics as follows:

- Safe – 1 vulnerability (LSP3-01): Focused on reducing human error by ensuring teachers are trained and retain oversight.
- Fair – with harmful bias managed – 1 vulnerability (LSP3-02): Addressing the risk of unequal treatment in personalized recommendations.
- Valid & Reliable – 1 vulnerability (LSP3-03): Concerned with sustaining model accuracy through ongoing monitoring and adaptation.

*4.5.4. LSP4 - Underwater Drones*

LSP4 explores the trustworthiness of AI systems used for autonomous inspection and maintenance assessment of maritime infrastructure. The deployment of AI-driven UAVs introduces a range of vulnerabilities related to safety, data integrity, transparency, and operational reliability. Table 14 summarizes the twelve unique vulnerabilities and related control measures as documented within LSP4, each mapped to its corresponding threat, AI asset, and trustworthiness (TWA) dimension/characteristic (the completed KB enhancement template can be found in Supplementary Figure 32, Annex C4). These include a wide range of technical and operational issues, from biased datasets and unsafe model outputs to insecure artefacts and inadequate process logging. The associated control measures emphasize both technical (e.g. encryption, geofencing, dataset augmentation) and procedural safeguards (e.g. human-in-the-loop validation, structured logging), providing a concrete foundation for threat mitigation and trustworthiness enhancement.

**Table 14.** The identified unique vulnerabilities, related control measures, corresponding threats, affected AI assets, and trustworthiness (TWA) dimension/characteristics for LSP4.

| ID | Associated Vulnerabilities | Associated control measures | Threat title | Affected AI Asset | Affected AI lifecycle stage | Affected TWA dimension/characteristic |
|---|---|---|---|---|---|---|
| LSP4-01 | Insufficient representation of defect types in validation dataset | 1. Expand defect dataset with real and synthetic samples | Misclassification of Structural Defects & Objects | Models | Testing | Valid & Reliable |
| LSP4-02 | Biased training data lacking environmental diversity | 1. Augment dataset with varied visibility, turbidity, and depth scenarios | Dataset Bias from Clear Water Imagery | Data | Design | Fair - with harmful bias managed |
| LSP4-03 | Lack of safety constraints in route generation | 1. Integrate geofencing and environmental risk maps into path planning logic | Unsafe Route Near Sensitive Structures | Models | Design | Safe |
| LSP4-04 | Unprotected mission configuration files and weak access controls | 1. Digitally sign and hash configuration files | Tampered Mission Parameters | Environment/tools | Production | Secure & Resilient |
| LSP4-05 | Lack of rationale generation in optimization logic | 1. Log scoring function output | Opaque Coverage Decisions | Models | Design | Explainable & Interpretable |
| LSP4-06 | Coverage optimization tuned on simple geometric environments | 1. Use synthetic complex structures in | Inspection Neglect Due to Biased Path | Models | Testing | Fair - with harmful bias managed |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | test maps | Sampling | | | |
| LSP4-07 | Missing telemetry and path comparison logs | 1. Store mission plan, executed path, and timestamps in a secure, queryable format | No Audit Trail for Route Execution | Processes | Production | Accountable & Transparent |
| LSP4-08 | Lack of anonymization or access control for inspection video datasets | 1. Apply encryption, role-based access, and redact sensitive visual elements before storage or sharing | Privacy Leakage via Inspection Footage | Data | Production | Privacy-enhanced |
| LSP4-09 | Obsolete or poorly integrated bathymetric and infrastructure maps | 1. Update environmental models regularly | Coverage Failure Due to Invalid Terrain Model | Data | Testing | Valid & Reliable |
| LSP4-10 | Insufficient protection of route logs and mission telemetry files | 1. Store logs in encrypted format | Route Logging Exposes Sensitive Infrastructure Layouts | Artefacts | Production | Privacy-enhanced |
| LSP4-11 | CNN lacks real-time environmental validation cross-check | 1. Introduce human-in-the-loop validation before diver deployment | Unsafe Classification Leading to Diver Deployment | Processes | Production | Safe |
| LSP4-12 | Route planning does not adapt to real-time oceanographic inputs | 1. Implement adaptive path re-planning using real-time current sensors and environment models | Suboptimal Route Execution Under Uncertain Currents | Models | Production | Valid & Reliable |

The unique vulnerabilities are distributed across the seven TWA dimensions/characteristics as follows:

- Valid & Reliable – 3 vulnerabilities (LSP4-01, LSP4-09, LSP4-12): These focus on model accuracy and environmental model validity, highlighting concerns about misclassification and failure under unpredictable conditions.
- Fair – with harmful bias managed – 2 vulnerabilities (LSP4-02, LSP4-06): These issues reflect biases in training and optimization data, particularly due to environmental homogeneity in simulation inputs.
- Safe – 2 vulnerabilities (LSP4-03, LSP4-11): These relate to route safety and model validation, with potential physical risks to infrastructure or human operators.
- Privacy-enhanced – 2 vulnerabilities (LSP4-08, LSP4-10): Privacy risks stem from unprotected video data and telemetry logs, raising concerns about unauthorized exposure of sensitive information.
- Secure & Resilient – 1 vulnerability (LSP4-04): Focused on protecting mission-critical configuration files from tampering.
- Explainable & Interpretable – 1 vulnerability (LSP4-05): Addressing the need for rationale and traceability in optimization outcomes.
- Accountable & Transparent – 1 vulnerability (LSP4-07): Highlighting the absence of auditing capabilities for route execution decisions.

The unique vulnerabilities are distributed across the AI-related assets as follows:

- Models – 5 vulnerabilities (LSP4-01, LSP4-03, LSP4-05, LSP4-06, LSP4-12): These include issues with classification, route generation, and optimization explainability.
- Data – 3 vulnerabilities (LSP4-02, LSP4-08, LSP4-09): Concerns relate to dataset diversity, privacy exposure, and outdated environmental maps.
- Processes – 2 vulnerabilities (LSP4-07, LSP4-11): These affect route auditing and the real-time validation of model outputs.
- Environment/Tools – 1 vulnerability (LSP4-04): Highlighting tampering risk in mission setup due to inadequate file protection.
- Artefacts – 1 vulnerability (LSP4-10): Related to the unencrypted storage of mission logs and telemetry files.

The unique vulnerabilities are distributed across the AI-related lifecycle stages as follows:

- Design – 3 vulnerabilities (LSP4-02, LSP4-03, LSP4-05): These involve early-stage issues such as biased training data, unsafe route planning logic, and lack of explainability in optimization,
- Testing – 3 vulnerabilities (LSP4-01, LSP4-06, LSP4-09): These highlight limitations in dataset coverage, simulation environment diversity, and outdated environmental models, and
- Production – 6 vulnerabilities (LSP4-04, LSP4-07, LSP4-08, LSP4-10, LSP4-11, LSP4-12): These concern runtime risks including tampering of mission configurations, inadequate logging, privacy leaks, and failures to adapt to real-time environmental conditions.

#### *4.5.5. LSP5 - Wastewater management*

LSP5 evaluates the trustworthiness of AI systems designed to support optimization in wastewater processing—specifically for regulating nitrogen levels. The deployment of predictive models in this context introduces a mix of technical and human-centric challenges. The vulnerabilities span explainability, fairness, accountability, and safety, especially considering that decisions informed by AI affect environmental quality, stakeholder responsibility, and operator trust. Table 15 summarizes the twelve unique vulnerabilities and related control measures as documented within LSP5, each mapped to its corresponding threat, AI asset, and trustworthiness (TWA) dimension/characteristic (the completed KB enhancement template can be found in Supplementary Figure 33, Annex C5). These vulnerabilities underscore not only technical shortcomings but also sociotechnical risks—such as

operator deskilling and over-reliance on automation—that can compromise trust and safe deployment.

**Table 15.** The identified unique vulnerabilities, related control measures, corresponding threats, affected AI assets, and trustworthiness (TWA) dimension/characteristics for LSP5.

| ID | Associated Vulnerabilities | Associated control measures | Threat title | Affected AI Asset | Affected AI lifecycle stage | Affected TWA dimension/characteristic |
|---|---|---|---|---|---|---|
| LSP5-01 | De-skilling of operators | 1. Transparent models, assessed with experienced operators. 2. Insight into complex relationship between factors | Deskilling from lack of transparency | Models | Testing | Accountable & transparent |
| LSP5-02 | Diffusion of stakeholder accountability | 1. Accountability strengthened through transparency and interpretability, evaluated with stakeholders | Diffusion of accountability due to lack of transparency | Models | Design, Testing | Accountable & Transparent |
| LSP5-03 | Unfair for different operator groups, e,.g. due to lack of universal design | 1. Human-centred evaluation of AI prediction presentations, e.g. as these are shown in the user interface / dashboard | Accessible presentation of AI model predictions | Artefacts | Design | Fair-with harmful bias managed |
| LSP5-04 | Operator responsibility requires explainability | 1. Human-centred evaluation with operators | Reduced operator accountability due to lack of interpretability | Models | Design, Testing | Explainable & Interpretable |
| LSP5-05 | Loss of operator trust | 1. Human-centred evaluation with operators | Reduced ability of operators to identify prediction errors | Models | Design, Testing | Explainable & Interpretable |
| LSP5-06 | Over-reliance in model predictions | 1. Human-centred evaluation with operators. 2. Provide training and AI skill development to stakeholders, including operators. | Over-reliance in model predictions | Models | Design, Testing | Explainable & Interpretable |
| LSP5-07 | Security vulnerabilities, | 1. Security control measures, as | Security threats - as | Models | Design, Testing | Secure & Resilient |

| | | | | | | |
|---|---|---|---|---|---|---|
| | as discussed in TrustGuard assessment | discussed in TrustGuard assessment | discussed in TrustGuard assessment | | | |
| LSP5-08 | Potential environmental risk from erroneous AI model predictions, leading to insufficient wastewater treatment | 1. Risk is low as the AI prediction models are planned to be deployed in an existing process with continuous monitoring of process output, and where the established control system will serve as fallback in case of issues. | Environmental risk from AI model error | Models | Design | Safe |

The unique vulnerabilities are distributed across the AI-related lifecycle stages as follows:

- Design – 6 vulnerabilities (LSP5-02, LSP5-03, LSP5-04, LSP5-05, LSP5-06, LSP5-08): These emphasize early design concerns such as stakeholder accountability, operator trust, explainability, and safety of environmental outcomes.
- Testing – 6 vulnerabilities (LSP5-01, LSP5-02, LSP5-04, LSP5-05, LSP5-06, LSP5-07): These highlight the need to evaluate human-AI interaction, transparency, and robustness during system validation phases.

The unique vulnerabilities are distributed across the seven TWA dimensions/characteristics as follows:

- Accountable & Transparent – 2 vulnerabilities (LSP5-01, LSP5-02): These point to reduced stakeholder clarity and operator role erosion due to limited transparency.
- Fair – with harmful bias managed – 1 vulnerability (LSP5-03): Highlights interface design gaps affecting operator inclusivity.
- Explainable & Interpretable – 3 vulnerabilities (LSP5-04, LSP5-05, LSP5-06): These focus on enhancing operator understanding and trust in model outputs.
- Secure & Resilient – 1 vulnerability (LSP5-07): Reflects system integrity and data protection needs during model development and testing.
- Safe – 1 vulnerability (LSP5-08): Emphasizes environmental safeguards and fallback mechanisms against prediction errors.

The unique vulnerabilities are distributed across the AI-related assets as follows:

- Models – 7 vulnerabilities (LSP5-01, LSP5-02, LSP5-04, LSP5-05, LSP5-06, LSP5-07, LSP5-08): These reflect risks in model transparency, security, safety, and their impact on operator roles.
- Artefacts – 1 vulnerability (LSP5-03): Addressing fairness in AI prediction interfaces and accessibility of system outputs.

#### *4.5.6. LSP6 - Healthcare*

LSP6 focuses on assessing the trustworthiness of an AI-based system designed to support radiologists in segmenting the prostate gland and its substructures from medical imaging data. This segmentation task is clinically significant and complex, requiring high accuracy, reliability, and adaptability to

heterogeneous clinical environments. Table 16 summarizes the eleven unique vulnerabilities and related control measures as documented within LSP6, each mapped to its corresponding threat, AI asset, and trustworthiness (TWA) dimension/characteristic (the completed KB enhancement template can be found in Supplementary Figure 34, Annex C6). These vulnerabilities underscore not only technical shortcomings—such as data drift, limited generalizability across institutions, and insufficient adversarial robustness—but also sociotechnical risks that can compromise trust in clinical deployment. For example, the lack of usability testing for explanation tools may erode radiologists' confidence in AI-assisted diagnostics, while fairness and privacy trade-offs introduced through over-anonymized datasets may inhibit both equity and transparency in model performance.

**Table 16.** The identified unique vulnerabilities, related control measures, corresponding threats, affected AI assets, and trustworthiness (TWA) dimension/characteristics for LSP6.

| ID | Associated Vulnerabilities | Associated control measures | Threat title | Affected AI Asset | Affected AI lifecycle stage | Affected TWA dimension/characteristic |
|---|---|---|---|---|---|---|
| LSP6-01 | The AI model lacks to perform correctly on newly acquired and processed data. The AI model experiences some performance drifts and lacks to produce correct results on data processed in production settings. Lack of performance monitoring tools able to detect model performance | 1. Adopt assessment tools to detect possible malfunctioning of the system (e.g., uncertainty quantification strategies, outliers detection, drift detection tools) and inform the user when there can be potential misbehaviour of the system | Lack of generalizability in production settings | Models | Production | Valid & reliable |
| LSP6-02 | The AI model was trained on images from just one or few hospitals or imaging devices and hence does not yield reliable results when applied to images from different sources or equipments. | 1. Use multicentre and multidevice datasets and assess all the possible variance sources of the data | Lack of generalizability in testing | Models | Testing | Valid & reliable |
| LSP6-03 | The AI model has been trained on datasets limited in scope, e.g., on data acquired from just one clinical centre or on data acquired with just one device. The AI model lacks to perform as expected on unseen data. | 1. Use multicentre and multidevice datasets and assess all the possible variance sources of the data | Lack of extensive data for training and testing | Models | Design, Testing | Valid & reliable |

| | | | | | | |
|---|---|---|---|---|---|---|
| LSP6-04 | The AI model is not able to work on local premises, but it requires external resources on the cloud. This may expose the system to online attacks. | 1. Adopt pricacy preserving strategies and ensure that the data transferred are fully anonimised before sharing on the Net or reduce the computation requirements to ensure the system can work on local premises also in case of low-resource settings | Need of external computation resources to infer segmentation | Models | Production | Privacy-enhanced |
| LSP6-05 | The AI model is not equipped with information about its working conditions, performance metrics and data used during training, testing and validation. | Adopt tools and strategies to document and track all the choices done and the data used | Lack of a complete documentation on AI model functioning | Models | Design | Accountable & transparent |
| LSP6-06 | The explanation functionalities have not been assessed for usability with radiologists and end-users | 1. Run user studies to understand expectations and requirements of radiologists and their preferences in terms of explanations of results | Lack of effective explanation methods | Models | Design | Explainable & Interpretable |
| LSP6-07 | The data on which the model is used are of poor quality. Poor design choices on data selection for training the model, which results in poor performances and generalization capabilities | 1. Adopt data-quality assessment tools to evaluate the data used during the training and testing phases | Poor quality of data used in training and testing | Data | Production | Valid & reliable |
| LSP6-08 | The model may perform poorly on patient populations not well-represented in the training data, leading to inaccurate | 1. Adopt tools to assess data skewness on various critical variables, such as ethnicity, age, | Limited diversity in training data | Data | Design | Fair-with harmful bias managed |

| | | | | | | |
|---|---|---|---|---|---|---|
| | segmentation in certain groups | pathological conditions. 2. Map all the sources of variability in the data and check them for your datasets. | | | | |
| LSP6-09 | The training and testing data are fully anonymised and lack relevant information that are needed to perform fairness analyses | 1. Adopt privacy preservation measures that enable keeping track of relevant information about sensible attributes in the clinical data | Lack of critical information in training and testing data | Data | Design, Testing | Privacy-enhanced, Fair-with harmful bias managed |
| LSP6-10 | The training and testing data are fully anonymised and lack relevant information thus hindering the correct functioning of the explanation tools | 1. Adopt privacy preservation measures that enable keeping track of relevant information about sensible attributes in the clinical data | Lack of critical information in training and testing data | Data | Design | Privacy-enhanced, Explainable & Interpretable |
| LSP6-11 | The model was trained without testing for its resilience to adversarial data | 1. Adopt assessment tools to detect possible mulfunctionings of the system (e.g., uncertainty quantification strategies, outliers detection, drift detection tools) and inform the user when there can be potential misbehaviour of the system. 2. Adopt adversarial training strategies to make the model resilient to corrupted or noisy data. | Lack of robustness checks | Models | Design | Secure & Resilient |

The unique vulnerabilities for LSP6 are distributed across the AI-related lifecycle stages as follows:

- Design – 7 vulnerabilities (LSP6-03, LSP6-05, LSP6-06, LSP6-08, LSP6-09, LSP6-10, LSP6-11): These emphasize the importance of data diversity, model documentation, usability of explanations, fairness assessments, and adversarial robustness considerations during the early stages of system development.
- Testing – 3 vulnerabilities (LSP6-02, LSP6-03, LSP6-09): These reflect the need to evaluate model generalizability, data adequacy, and fairness analysis during system validation phases.
- Production – 3 vulnerabilities (LSP6-01, LSP6-04, LSP6-07): These highlight operational concerns such as drift detection, dependency on external infrastructure, and data quality assurance in real-world deployment settings.

The unique vulnerabilities are distributed across the seven TWA dimensions/characteristics as follows:

- Valid & Reliable – 5 vulnerabilities (LSP6-01, LSP6-02, LSP6-03, LSP6-07): These reflect concerns around data quality, model generalization, and monitoring across the AI lifecycle, especially in medical imaging contexts.
- Privacy-enhanced – 3 vulnerabilities (LSP6-04, LSP6-09, LSP6-10): These highlight the challenge of balancing data anonymization with usability for fairness and explanation.
- Fair – with harmful bias managed – 2 vulnerabilities (LSP6-08, LSP6-09): These reveal gaps in demographic representativeness and the preservation of fairness-related attributes in anonymized data.
- Explainable & Interpretable – 2 vulnerabilities (LSP6-06, LSP6-10): These relate to the usability of explanation methods and the preservation of necessary data attributes for interpretability.
- Accountable & Transparent – 1 vulnerability (LSP6-05): Reflects the need for comprehensive documentation and traceability of model development.
- Secure & Resilient – 1 vulnerability (LSP6-11): Emphasizes the importance of robustness and adversarial resilience testing in model design.

The unique vulnerabilities are distributed across the AI-related assets as follows:

- Models – 8 vulnerabilities (LSP6-01, LSP6-02, LSP6-03, LSP6-04, LSP6-05, LSP6-06, LSP6-11): These represent risks associated with performance, explainability, security, and documentation across the lifecycle.
- Data – 4 vulnerabilities (LSP6-07, LSP6-08, LSP6-09, LSP6-10): These indicate critical concerns around data quality, diversity, fairness-related attribute availability, and privacy preservation.

### *4.5.7. LSP7 - Active Aging*

LSP7 explores AI-driven monitoring systems designed to assess frailty progression and daily activity patterns in older adults. These systems rely on ambient sensors and machine learning for early anomaly detection and personalized care support. The vulnerabilities identified in this context concern issues of model calibration, drift sensitivity, transparency in data processing, and potential privacy breaches—especially in low-data or sensitive environments. Table 17 summarizes the twelve unique vulnerabilities and related control measures as documented within LSP7, each mapped to its corresponding threat, AI asset, and trustworthiness (TWA) dimension/characteristic (the completed KB enhancement template can be found in Supplementary Figure 35, Annex C7).

**Table 17.** The identified unique vulnerabilities, related control measures, corresponding threats, affected AI assets, and trustworthiness (TWA) dimension/characteristics for LSP7.

| ID | Associated Vulnerabilities | Associated control measures | Threat title | Affected AI Asset | Affected AI lifecycle stage | Affected TWA dimension/chara |
|---|---|---|---|---|---|---|

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | **cteristic** |
| LSP7-01 | Excessive false positive alerts due to uncalibrated thresholds | 1. Define and monitor indicators for proper functioning of the model | Failure or malfunction of ML application | Models | Production | Valid & Reliable |
| LSP7-02 | Semantic drift in cluster definitions | 1. Define and monitor indicators for proper functioning of the model | Failure or malfunction of ML application | Models | Production | Valid & Reliable |
| LSP7-03 | Sensitivity to drift under low data volume | 1. Include or integrate in the AI model a drift control service | Failure or malfunction of ML application | Models | Production | Valid & Reliable |
| LSP7-04 | Variability in sensor installations and non-transparent EDA processes | 1. Ensure that the model is sufficiently resilient to the environment in which it will operate | Failure or malfunction of ML application | Data | Design | Explainable & Interpretable |
| LSP7-05 | Lack of occupant discrimination: no mechanism to tell the primary user's IR signature apart from other people or heat sources in the room | 1. Ensure that the model is sufficiently resilient to the environment in which it will operate | Failure or malfunction of ML application | Models | Production | Valid & Reliable |
| LSP7-06 | Data implicitly encodes location: no explicit sensor-to-room mapping | 1. Ensure that the model is sufficiently resilient to the environment in which it will operate | Failure or malfunction of ML application | Data | Production | Valid & Reliable |
| LSP7-07 | Reconstruction of personal information from aggregated data under low user count | 1. Ensure that models respect differential privacy | Data disclosure | Data | Production | Privacy-enhanced |
| LSP7-08 | No anomalies detected (zero recall) | 1. Define and monitor indicators for proper functioning of the model | Failure or malfunction of ML application | Models | Production | Valid & Reliable |

The unique vulnerabilities are distributed across the AI-related lifecycle stages as follows:

- Design – 1 vulnerability (LSP7-04): Addressing the transparency and variability in early-stage data analysis and system deployment environments.
- Production – 7 vulnerabilities (LSP7-01, LSP7-02, LSP7-03, LSP7-05, LSP7-06, LSP7-07, LSP7-08): These concern operational reliability, sensitivity to drift, misclassification risks, and privacy challenges under real-world conditions.

The unique vulnerabilities are distributed across the AI-related assets as follows:

- Models – 5 vulnerabilities (LSP7-01, LSP7-02, LSP7-03, LSP7-05, LSP7-08): Primarily targeting model calibration, detection accuracy, and resilience to change.
- Data – 3 vulnerabilities (LSP7-04, LSP7-06, LSP7-07): Addressing data handling risks including transparency, mapping issues, and privacy breaches.

The unique vulnerabilities are distributed across the seven TWA dimensions/characteristics as follows:

- Valid & Reliable – 6 vulnerabilities (LSP7-01, LSP7-02, LSP7-03, LSP7-05, LSP7-06, LSP7-08): Highlighting the need for dependable operation, especially under low-data and high-variability conditions.
- Explainable & Interpretable – 1 vulnerability (LSP7-04): Targeting transparency in exploratory data processing and adaptability to deployment setups.
- Privacy-enhanced – 1 vulnerability (LSP7-07): Focused on limiting re-identification risks in low-population data through privacy-preserving techniques.

# 5. Conclusions and Future Steps

The stakeholder survey conducted as part of the FAITH project has provided valuable empirical insights into current practices, challenges, and perceptions surrounding trustworthy and human-acceptable AI. Respondents from a wide range of organizational and sectoral backgrounds reported engagement with diverse technical and organizational measures—such as bias mitigation strategies, explainability tools, data anonymization, and compliance with standards like GDPR, ISO, and the AI Act. However, there remains significant heterogeneity in both the maturity and consistency of these approaches, with some respondents indicating a lack of formal governance structures or uncertainty regarding applicable regulations. A particularly insightful dimension of the study emerged from the Bayesian analysis of stakeholder understanding and acceptance of AI. This analysis provided probabilistic modelling of patterns in participant responses, enabling a more nuanced interpretation of their perceptions and awareness. The results suggest that while a majority of stakeholders exhibit a positive disposition toward the adoption of AI, this acceptance is conditional and context-dependent. Acceptance levels were significantly higher among those who also demonstrated a greater understanding of AI's functionality, limitations, and ethical implications—highlighting a critical link between knowledge and trust.

Conversely, the Bayesian inference revealed that lower levels of AI understanding—particularly regarding the principles of explainability, bias, and legal accountability—correlated strongly with hesitation, skepticism, or passive acceptance of AI systems. This suggests that trust in AI is not merely a product of exposure or familiarity, but is fundamentally shaped by the depth of stakeholder comprehension. Furthermore, the analysis showed that participants with better awareness of regulatory frameworks (such as the GDPR and AI Act) were more likely to express confidence in the systems their organizations deploy, reinforcing the importance of legal-literacy and ethical alignment in building trust. These findings have critical implications for the future development and operationalization of the FAITH Trustworthiness Assessment Framework (FAITH AI_TAF). First, they underscore the need for education and communication tools tailored to different stakeholder groups, especially those with lower baseline understanding of AI. Second, they point to the necessity of embedding trust-building features—such as transparency-by-design and user-centered explanations—directly into the technical architecture of AI systems, rather than treating them as post hoc compliance steps.

In the upcoming phases, FAITH will use these insights to enhance the design of co-creation workshops, interactive guidance materials, and domain-specific trustworthiness toolkits. These activities will explicitly focus on bridging gaps between perception and implementation, promoting not just technically robust AI, but systems that are understandable, acceptable, and aligned with human values. Moreover, the Bayesian findings will inform the development of dynamic user profiling capabilities within the FAITH tools, enabling assessments that adapt to the stakeholder's trust readiness and literacy. In parallel, FAITH will disseminate the results through scientific publications and community engagement, aiming to foster a broader dialogue around trustworthy AI that is grounded in both quantitative evidence and participatory feedback. By doing so, the project contributes not only to methodological advancement but also to the societal uptake of AI systems that are transparent, fair, and deserve public confidence.

In addition to the survey analysis, the current deliverable presents a thorough collection of stakeholder-driven requirements for trustworthy and human-acceptable AI. These were gathered through structured interviews, co-design workshops, and iterative engagement with each LSP, enabling the identification of practical challenges, expectations, and domain-specific needs. The analysis was complemented by an in-depth mapping of the applicable legal and regulatory regimes for each pilot site (Section 2), covering instruments such as the GDPR, AI Act, and sector-specific

legislation (e.g., MDR, DSA). This mapping provided essential context for interpreting stakeholder input and aligning it with normative obligations. Importantly, the collected requirements and legal insights were translated into actionable inputs for enhancing the FAITH KB. Through this iterative process, each LSP contributed to identifying vulnerabilities, associated threats, relevant trustworthiness dimensions, and appropriate mitigation measures. The FAITH KB is a living and evolving database which serves as a central component to guide the design, assessment, and communication of AI trustworthiness throughout the FAITH ecosystem.

Overall, the activities carried out under Task 2.3, as presented in this Deliverable D2.5, significantly contribute to various core areas of the FAITH project. They play a central role in the continuous evolution of the FAITH AI_TAF by integrating feedback collected through co-design workshops and stakeholder surveys. This input enables the framework to be refined in line with the legal, ethical, and technical requirements expressed by the LSPs. Furthermore, Task 2.3 actively supports the collaborative design and development of the FAITH Knowledge Base (KB), incorporating empirical insights on trustworthiness, contextual risk factors, and end-user perspectives. This reinforces the KB's role as a collaborative and evolving knowledge-sharing tool. The analysis of the initial anonymous survey serves as a foundational reference point for understanding stakeholder perceptions of trustworthy AI. In addition, the outcomes of Task 2.3 contribute to regulatory and ethical alignment by mapping stakeholder expectations against the legal and ethical risk areas addressed in WP1. Ultimately, trustworthy AI cannot be reduced to technical compliance alone. Trust is a dynamic and inherently human process, shaped by transparency, shared understanding, and meaningful engagement. As such, FAITH shall remain responsive to societal expectations, stakeholder experiences, and evolving legal standards. Building and sustaining trust in AI requires continuous reflection, inclusive co-creation, and a commitment to aligning technological development with human values.

# 6. References

In the current deliverable the references are denoted as footnotes.

# Appendix

A. The FAITH 1st Anonymous Survey



FAITH 1st anonymous survey

Fields marked with * are mandatory.

**FAITH - Fostering Artificial Intelligence Trust for Humans towards the optimization of trustworthiness through large-scale pilots in critical domains**

**FAITH 1st Anonymous survey**

*Purpose of this survey*

*FAITH aims to provide the practitioners and stakeholders of AI systems not only with a comprehensive analysis of the foundations of AI trustworthiness but also with an operational playbook for how to assess, and build trustworthy AI systems and how to continuously measure their trustworthiness.*

*Objective*

*The objective of this survey is to gain a deeper insight into the current practices and attitudes towards responsible and trustworthy AI. To provide context, this survey is intended for data scientists, AI engineers, and developers. It has been developed within the EU HE FAITH project.*

*With regard to the information that will be collected, in addition to some preliminary questions concerning your profile, the remaining questions pertain to the strategies employed in relation to the core dimensions of AI trustworthiness. Some questions are highly specialised and it is not anticipated that you have experience in all of them.*

*When completing the questionnaire (especially for the open-ended questions), please refrain from reporting any personal data either yours or others. According to GDPR, 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.*

*All responses will remain anonymous. No email or IP address will be recorded unless you opt to provide your email address, so that we can notify you of the results of this survey. Prior to distribution, the questionnaire was reviewed and approved by the Research Ethics Committee of FORTH.*

*The FAITH project will have access to the results of the survey for two purposes:*

*a) to map and report on the current state of play as regards the strategies for trustworthy and responsible AI,b) to use the results to provide a framework integrating the most commonly used tools.*

*A paper based on the results of this study might be prepared for publication; it will present the findings in a summary format.*

1

*Please note that the raw data **will not be made available in an open format**, but will remain under the control of the FAITH project (Responsible Partner: FORTH).*

***Confidentiality policy***

*Foundation for Research and Technology – Hellas (FORTH) is compliant with all legal procedures in respect of personal data processing, as set out in the applicable national (Act 4624/2019) and European law, and the EU Regulation on the protection of natural persons with regard to the processing of personal data (EU/2016/679).*

*The information that will be collected through this questionnaire is anonymous, which means that it does not correspond to an identified or identifiable natural person.*

*While replying to the questions below (especially open questions) please avoid to mention any personal information, either yours or others.*

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

*Any publication of the results of this survey will be done in anonymous form to inform the public and the scientific community.*

***Notice***

*It is important to clarify that you cannot withdraw the provided answers to the questionnaire since it is conducted anonymously.*

***Communication***

*Further information on this survey can be obtained from the scientific responsible of this project. Your complaints shall be submitted to the scientific responsible of this project.*

*Contact details: **Prof. D.I Fotiadis, FORTH, email: fotiadis@uoi.gr.***

***Contact***

If you have any questions relating to personal data protection policy or if you believe that there is any vagueness or incompatibility in your personal data protection policy, you may communicate such instance to the **Data Protection Officer (DPO) of FORTH** by sending an e-mail at **dpo@admin.forth.gr**.

***Related Glossary of Terms***

**Fundamental Rights:** These are basic rights and freedoms that are recognized by national and international laws. In the European context, fundamental rights are guaranteed by instruments like the EU Charter of Fundamental Rights and include rights such as the right to privacy, freedom of expression, non-discrimination, and the right to a fair trial. In the context of AI, it is important to ensure that AI systems do not violate these rights, for example by making unfair decisions, infringing on privacy, or discriminating against individuals or groups. Key points:

- Encompasses rights to privacy, non-discrimination, and freedom of expression.
- AI systems should respect and uphold these rights in their design and deployment.
- Violations can lead to significant legal and ethical consequences.

**GDPR (General Data Protection Regulation):** A regulation in EU law focuses on data protection and privacy for all individuals within the European Union (EU) and the European Economic Area (EEA). It sets

2

strict guidelines on how personal data must be collected, stored, and used, ensuring individuals' control over their personal information. Non-compliance can result in heavy fines. Key points:

- Protects individuals' data privacy.
- Requires clear consent for data processing.
- Mandates that organizations inform users of any data breaches.

**NIS2 Directive (Network and Information Security Directive 2):** An updated European Union directive aimed at enhancing cybersecurity across member states, particularly for organizations involved in critical sectors (e.g., healthcare, finance, energy). It imposes stricter cybersecurity measures, incident reporting obligations, and a framework for cooperation between EU states on cybersecurity threats. Key points:

- Applicable to essential service providers.
- Strengthens cybersecurity requirements.
- Requires quick reporting of cybersecurity incidents.

**Anti-Discrimination Laws:** These are laws that prevent discrimination on the basis of race, gender, disability, sexual orientation, or other protected characteristics. In the AI context, these laws apply to ensuring that AI models do not produce biased outcomes that could result in unfair treatment of individuals or groups. Key points:

- Protects individuals from unfair treatment in employment, housing, and other sectors.
- In AI, refers to ensuring fairness and eliminating biases in algorithms and decision-making processes.

**AI Act (Artificial Intelligence Act):** A proposed legal framework by the European Union aimed at regulating the use and deployment of artificial intelligence. It establishes rules to ensure that AI systems are trustworthy and safe, with a focus on high-risk AI applications (e.g., in healthcare, transportation). The AI Act will categorize AI systems based on the level of risk they pose. Key points:

- Classifies AI applications into different risk categories (e.g., minimal, high-risk).
- Imposes strict regulations for high-risk AI systems (e.g., medical devices, autonomous vehicles).
- Aims to protect fundamental rights and prevent harm.

**Sector-Specific Regulations:** Laws and standards that apply specifically to certain industries, such as healthcare, transportation, or energy. These regulations ensure compliance with safety, data protection, and ethical standards within each sector. They aim to ensure sector compliance with safety, privacy, and ethical standards.

**Cybersecurity Laws:** A set of laws designed to protect the integrity, confidentiality, and availability of systems, networks, and data. These laws ensure that organizations have adequate security measures in place to prevent cyberattacks, data breaches, and unauthorized access to sensitive information. Key points:

- Enforces security standards for organizations to protect against cyber threats.
- Requires timely reporting of breaches.
- May include penalties for non-compliance with security standards.

3

**ISO Standards (International Organization for Standardization):** A set of internationally recognized standards that ensure quality, safety, efficiency, and consistency across products, services, and systems. In AI, specific standards such as ISO/IEC 38500 or ISO 27001 help organizations manage data security and governance effectively. Key points:

- Provides a globally recognized framework for quality and safety.
- In AI, helps ensure systems are designed and operated securely and reliably.
- Includes guidelines for managing information security risks and other risks concerning AI lifecycle.

**CEN/CENELEC (European Committee for Standardization and European Committee for Electrotechnical Standardization):** CEN and CENELEC are European standardization bodies that develop and implement technical standards across Europe, including in emerging fields such as AI. They aim to create harmonized standards that ensure AI systems are safe, secure, and aligned with European values and regulations. Their work is closely tied to the AI Act and other regulatory frameworks. Key points:

- Establishes European standards for AI safety, security, and performance.
- Focuses on harmonizing standards across Europe to ensure interoperability.
- Works alongside the EU to support regulatory compliance, including AI Act requirements.

**ETSI (European Telecommunications Standards Institute):** ETSI is an independent, not-for-profit standards organization that develops globally applicable standards for information and communication technologies (ICT). It also plays a key role in AI, focusing on AI's technical aspects, particularly in telecommunications, cybersecurity, and IoT (Internet of Things) applications. Key points:

- Sets standards for AI deployment in ICT and telecommunications.
- Works on AI-related security, privacy, and interoperability in networked environments.

☐ I hereby consent that I have read the Purpose, the Objective, and the Confidentialy policy of the survey

☐ I confirm that I am 18 years old or older

☐ I consent to voluntarily take part in this anonymous survey

☐ I have full legal capacity to answer this anonymous survey

## 1 General group of questions

* 1.1 Respondent's Profile (select one or more options)
  - ☐ Computer science or information engineering
  - ☐ Data science
  - ☐ Engineering
  - ☐ LSP domain experts
  - ☐ Mathematicians

4

* 1.2 Years of expertise in the domain? (select only one option)
  - 0-2
  - 2-5
  - 5-10
  - 11+

* 1.3 What type of organization do you work in? (select one or more options)
  - Tech Company
  - Public Organization
  - Research Centre
  - Private Company on a specific sector (please specify which one in the text below)

1.4 If you selected Private Company in the previous question, please provide the answer here

* 1.5 What approach do you adopt in AI developing process to ensure transparency? (select one or more options)
  - MLOps
  - Precise Documentation of Model Training
  - Model Card
  - Model Sharing on GitHub
  - Other (please specify which one in the text below)

1.6 If you selected Other in the previous question, please provide the answer here

* 1.7 What types of reliability metrics of the developed AI models in your domain do you utilize? (select one or more options)
  - Generalization Capacity on External Datasets
  - Sensitivity Analysis
  - Uncertainty Analysis and Quantification
  - Robustness to Adversarial Examples
  - Post-Deployment Performance Monitoring
  - Other (please specify which one in the text below)

1.8 If you selected Other in the previous question, please provide the answer here

* 1.9 What methods or practices does your organization use to analyze and detect harmful bias in AI-based systems? (select one or more options)
  - Pre-deployment Bias Testing
  - Post-deployment Monitoring
  - No current Practises

5

* 1.10 What are the ethical concerns you care the most during the AI lifecycle? (select one or more options)
  - ☐ Data Fairness
  - ☐ Model Fairness
  - ☐ Data Privacy
  - ☐ Data Security
  - ☐ System Safety
  - ☐ System and Model Security
  - ☐ System, Model Robustness and Generalization Capacity
  - ☐ System and Model Transparency
  - ☐ System Decision Opacity

* 1.11 How much time do you spend with the end-users/domain experts to define and test your AI system? (select only one option)
  - ○ Less than one month.
  - ○ Between 1 and 6 months.
  - ○ Between 6 and 12 months.
  - ○ More than one year

* 1.12 How do you secure the privacy threat of your input dataset? (select one or more options)
  - ☐ Data Anonymization Techniques
  - ☐ Federated Learning approach
  - ☐ Offline Learning
  - ☐ Synthetic Data Generation Techniques

* 1.13 Can you provide the above mentioned methods as open-source code? (select only one option)
  - ○ Yes
  - ○ No

* 1.14 How does your organization assess and communicate the risks associated with AI-systems to end-users and other interested stakeholders (e.g. domain experts, business leaders, security teams, quality assurance engineers)? (select one or more options)
  - ☐ Through detailed documentation and reporting
  - ☐ By holding regular briefings and training sessions
  - ☐ Via transparency tools like AI model cards or explainability features
  - ☐ Using risk assessment frameworks (e.g., FAITH AI_TAF, NIST AI RMF)
  - ☐ Other (please specify which one in the text below)

1.15 If you selected Other in the previous question, please provide the answer here

* 1.16 How often does your organization engage in co-creation activities with end-users and other interested stakeholders to align AI system development with their needs and concerns? (select one or more options)
  - ☐ Regularly and consistently throughout the project lifecycle.
  - ☐ At key stages of the development process (e.g., design, testing).
  - ☐ Occasionally, based on specific project needs.

6

- Rarely or never.
- Not applicable.

* 1.17 How does compliance with regulatory frameworks (e.g., GDPR, AI Act) influence your organization's AI system design and deployment strategies? (select only one option)
  - It has a significant influence, guiding all aspects of design and deployment.
  - It influences some parts of the process, particularly data handling.
  - It is considered mainly during the final stages of deployment.
  - Minimal influence; mainly an afterthought.
  - Not applicable.

* 1.18 What do you see as the most significant challenges in achieving trustworthy AI in your sector? (select one or more options)
  - Managing and mitigating biases in AI systems.
  - Ensuring data privacy and security.
  - Enhancing transparency and explainability of AI models.
  - Aligning AI systems with ethical and societal values.
  - Keeping up with regulatory changes and compliance requirements.

* 1.19 How does your organization measure success in addressing critical Trustworthy dimensions such as fairness, transparency, and robustness in your AI systems? (select one or more options)
  - Through predefined KPIs and performance metrics.
  - Via third-party audits and certifications.
  - By continuous monitoring and evaluation post-deployment.
  - Using user feedback and satisfaction surveys.

* 1.20 Approximately how much time does your team spend collaborating with domain experts during the AI model design and testing phases? (select only one option)
  - Less than one month.
  - 1-3 months
  - 3-6 months
  - 6-12 months
  - More than one year.

* 1.21 How does your organization ensure continuous improvement and adaptation of AI models to changing data and environments? (select one or more options)
  - Implementing a continuous feedback loop with users and stakeholders.
  - Regularly retraining models with new data to maintain accuracy.
  - Using automated monitoring systems to detect and address performance drifts.
  - Conducting periodic audits and assessments of model performance.
  - Engaging in collaborative research and development with external experts.

* 1.22 What ethical guidelines and governance frameworks does your organization follow to ensure responsible AI development and deployment? (select one or more options)
  - Adhering to international standards such as the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.

7

- [ ] Following guidelines from the European Commission's Ethics Guidelines for Trustworthy AI.
- [ ] Implementing internal ethics review boards or committees.
- [ ] Conducting regular ethics training for AI developers and stakeholders.
- [ ] Engaging with external ethicists and experts for independent reviews.
- [ ] None at the moment.

* 1.23 Do you believe that the FAITH AI model passport can be used for the monitoring of the AI systems in your organization? (select only one option)
  - ( ) Yes
  - ( ) No
  - ( ) I don't know

The FAITH AI model passport monitors, tracks and lists, in a standardised format (i.e. as metadata), all the model's details (e.g. models' scope, actors and developers involved, development tools and technical choices made, versions, performance metrics, and various trustworthiness characteristics), as well as, all the details of the data used to train it (e.g. data provenance and localization).

* 1.24 Do you believe that SpyderRisk can be useful for assessing AI based systems in your organization? (select only one option)
  - ( ) Yes
  - ( ) No
  - ( ) I don't know

Spyderisk is for assessing risk in complex socio-technical systems. By complex systems we mean "systems of systems that humans are unable to fully understand, debug or predict, typically socio-technical-physical systems." In other words, humans build critical systems that are unknowable, and so we need to apply mathematics to delineate the risks in these unknowable systems. Link: https://github.com/Spyderisk

* 1.25 How does your organization ensure that AI models are understandable to end-users and stakeholders, including providing explanations for decisions made by the AI? (select one or more options)
  - [ ] Via ad-hoc explainability methods applied to the AI models.
  - [ ] By selecting explainable-by-design methods.
  - [ ] By describing the processes followed to developed the AI models directly to involved stakeholders, but without any specific automated methods.
  - [ ] At the moment, no explainability methods are in use.

1.26 What measures does your organization take to assess and mitigate biases in AI systems to ensure fair treatment?

* 1.27 How does your organization test and validate AI systems to ensure they can handle a variety of inputs and operate reliably under different conditions? (select one or more options)
  - [ ] Sensitivity Analysis
  - [ ] Stability Analysis
  - [ ] Sandbox Testing

8

- [ ] Other (Please specify which one in the text below)

1.28 If you selected Other in the previous question, please provide the answer here

*1.29 How does your organization saleguard AI systems against malicious attacks and ensure the integrity and confidentiality of the data they process? (select one or more options)

- [ ] Adversarial Training and Testing
- [ ] Restricted Internet Access
- [ ] None

*1.30 What processes are in place within your organization to monitor AI system performance, address any issues that arise, and ensure accountability for the outcomes produced by AI? (select one or more options)

- [ ] In-house methods
- [ ] Gratana, Prometheus, ML flow, etc.
- [ ] None

*1.31 How does your organization ensure that AI systems consistently perform as intended and deliver accurate and dependable results over time? (select one or more options)

- [ ] Sandbox Testing
- [ ] Formal Verification Methods
- [ ] None

*1.32 Does your organization provide to the AI engineers guidelines and good practices for responsible AI development? (select only one option)

- ( ) Yes
- ( ) No
- ( ) I don't know

## 2 Domain specific group of questions

*2.1 What is the most critical Trustworthiness dimension from your working organization/pilot /domain? (select one or more options)

- [ ] Explainability and Interpretability
- [ ] Fairness
- [ ] Privacy Preservation
- [ ] Robustness and Reliability
- [ ] Safety
- [ ] Security
- [ ] Transparency and Accountability

*2.2 Have you implemented anything to address these critical dimensions, if it is quantifiable? (select only one option)

- ( ) Yes

9

- No

2.3 If you answered yes to the previous question, please specify the method you have implemented

* 2.4 Is your method open-access? (select only one option)
  - Yes
  - No

* 2.5 What are the relevant legislation, standards, policies (for the AI based systems) in your organisation?

* 2.6 Are your involved stakeholders in AI-based systems understand AI? (select from 1-5 scale, 1 being not at all)
  - 1
  - 2
  - 3
  - 4
  - 5

* 2.7 Are your involved stakeholders in AI-based systems accept the AI? (select from 1-5 scale, 1 being not at all)
  - 1
  - 2
  - 3
  - 4
  - 5

* 2.8 Identify the key considerations for trustworthy & human acceptable AI based systems in your organization

## 3 Group of legal and ethical questions

* 3.1 How often are legal and ethics experts involved in the development or deployment of AI systems? (select only one option)
  - Always involved
  - Frequently involved
  - Occasionally involved
  - When necessary
  - Rarely involved
  - Never involved

10

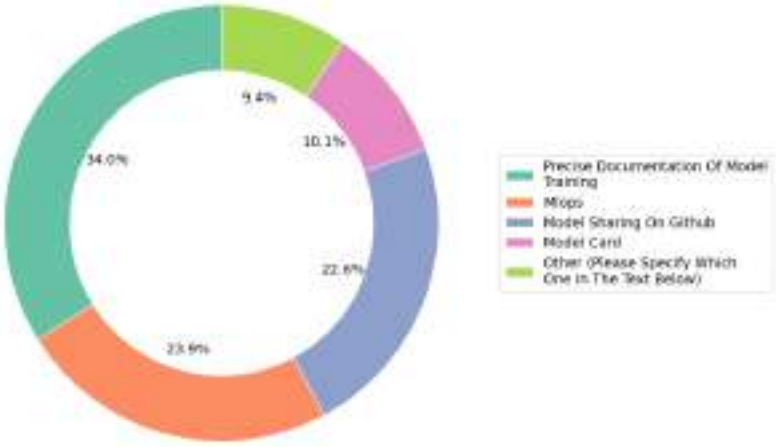
* 3.2 What is your opinion on the involvement of legal and ethics experts in the AI development and deployment process? (select only one option)
  - ○ Essential for ensuring compliance and accountability
  - ○ Important but could be more streamlined
  - ○ Useful, but not always necessary
  - ○ Involvement is minimal and should be increased
  - ○ Not required, AI development can proceed without it

* 3.3 What legal concerns and risks do you consider most important throughout the AI lifecycle? (select one or more options)
  - ☐ Violation of fundamental rights
  - ☐ Compliance with data protection and privacy laws (e.g., GDPR)
  - ☐ Fairness and non-discrimination in AI models
  - ☐ Compliance with anti-discrimination laws
  - ☐ Liability risks for system safety and failures
  - ☐ Compliance with cybersecurity laws (e.g., NIS2) for data and model security
  - ☐ Transparency and accountability for decision-making processes
  - ☐ Compliance with sector-specific regulations (e.g., healthcare, education, transportation) or standards (e.g., ISO, CEN/CENELEC, ETSI)

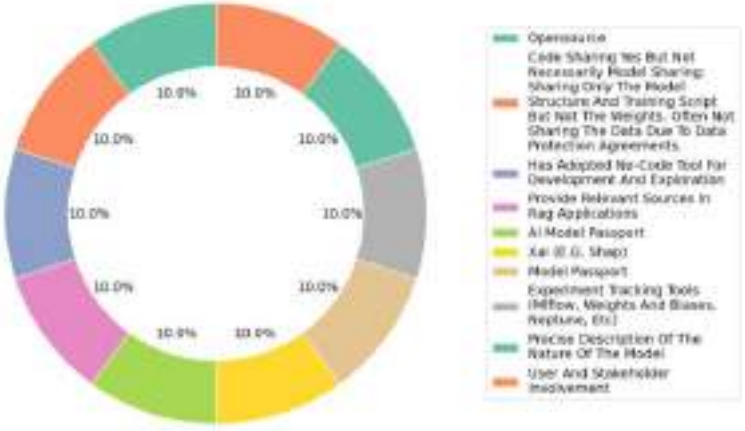
11

# B. Detailed survey analysis results

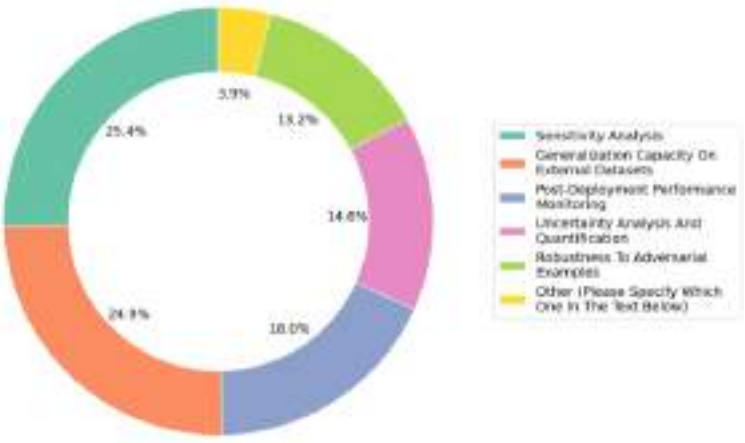
## B1. Overall



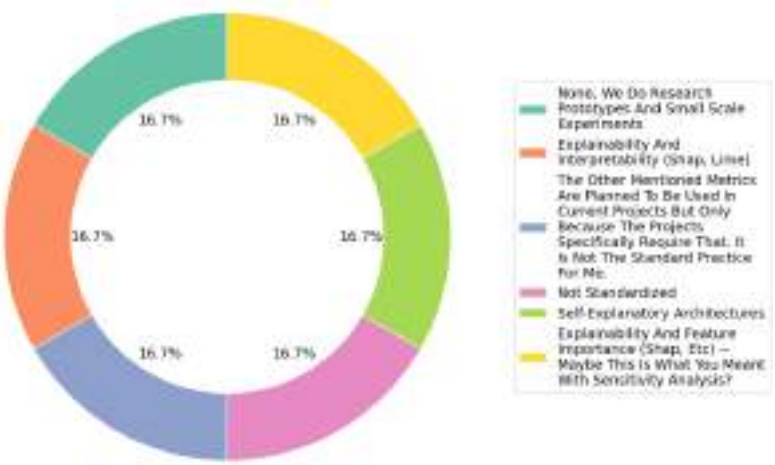

**Supplementary Figure 1.** Distribution plot of the responses from Q1.5 (Overall analysis).



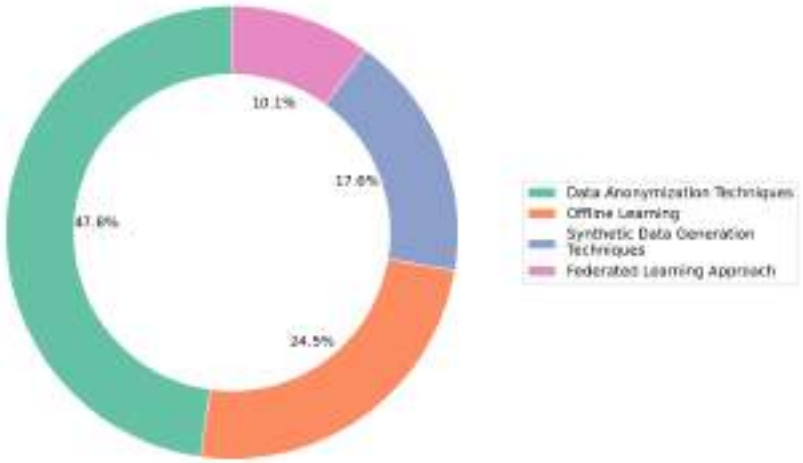

**Supplementary Figure 2.** Distribution plot of the responses from Q1.7 (Overall analysis).



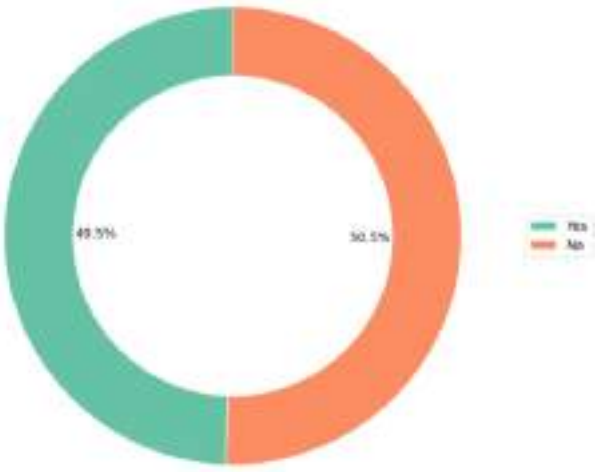

**Supplementary Figure 3.** Distribution plot of the responses from Q1.7 (Overall analysis).



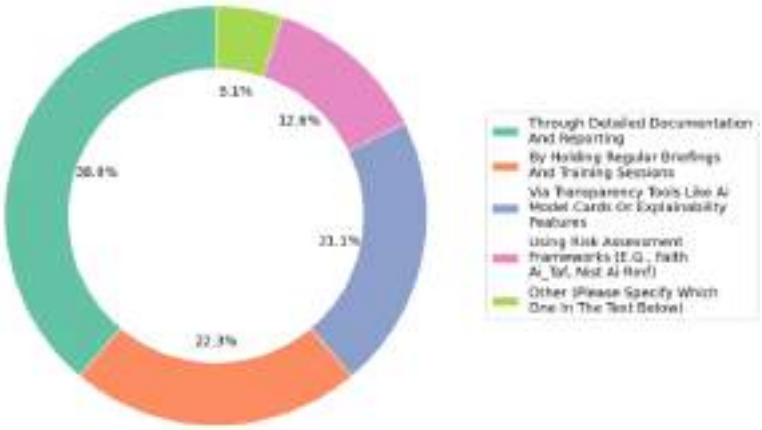

**Supplementary Figure 4.** Distribution plot of the responses from Q1.8 (Overall analysis).



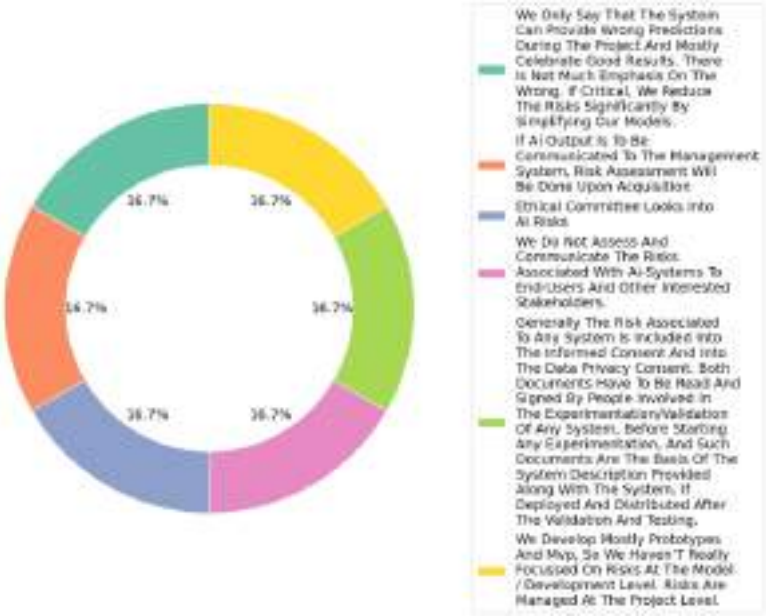

**Supplementary Figure 5.** Distribution plot of the responses from Q1.12 (Overall analysis).



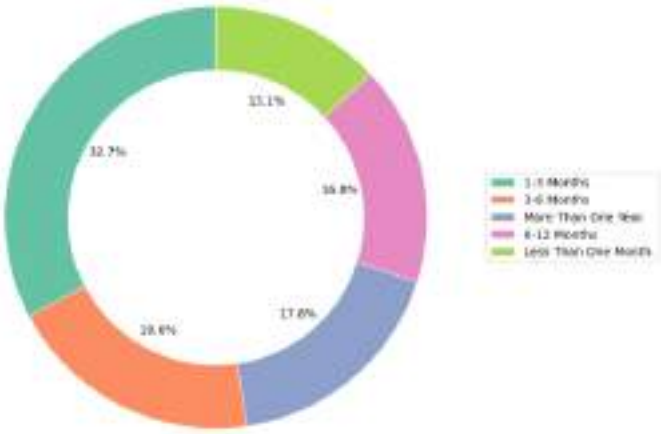

**Supplementary Figure 6.** Distribution plot of the responses from Q1.13 (Overall analysis).



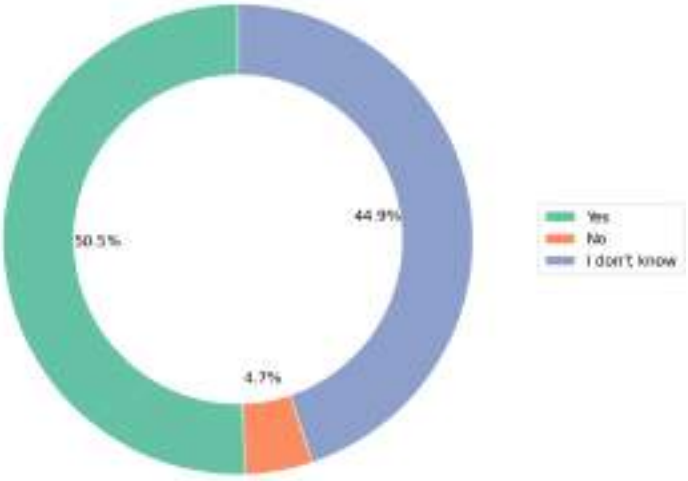

**Supplementary Figure 7.** Distribution plot of the responses from Q1.14 (Overall analysis).



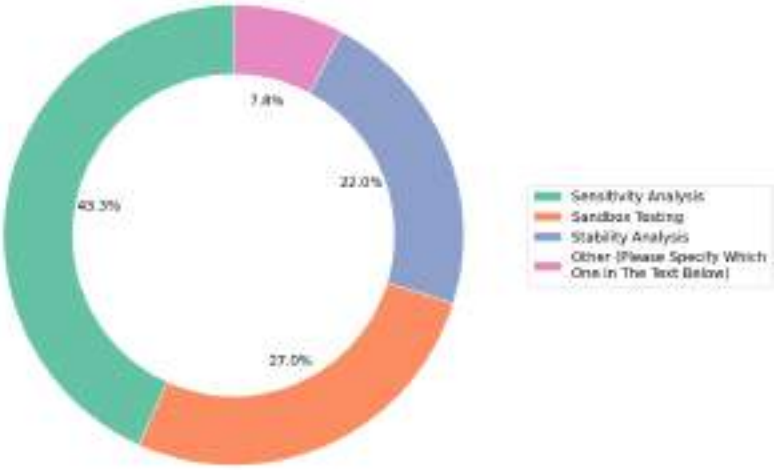

**Supplementary Figure 8.** Distribution plot of the responses from Q1.15 (Overall analysis).



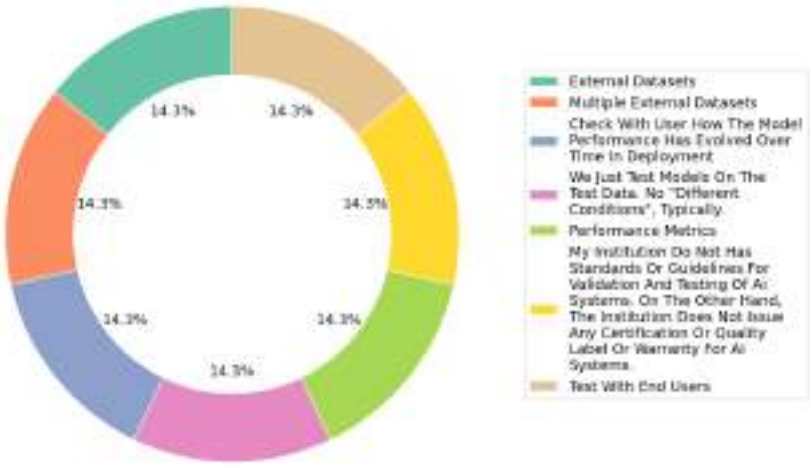

**Supplementary Figure 9.** Distribution plot of the responses from Q1.20 (Overall analysis).



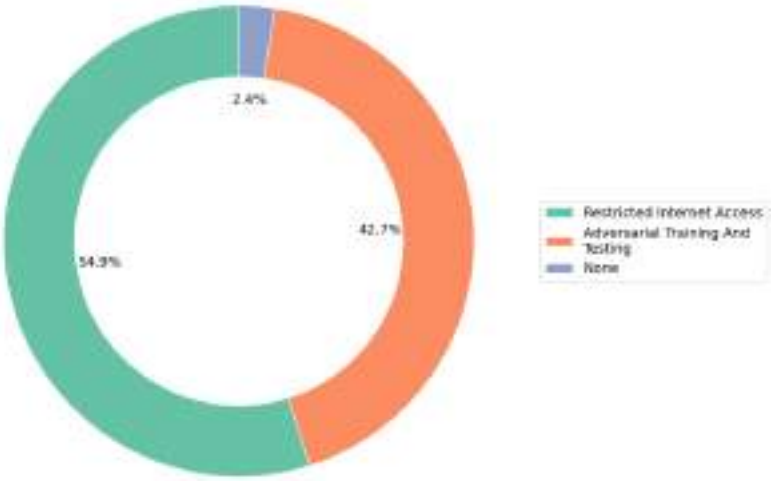

**Supplementary Figure 10.** Distribution plot of the responses from Q1.24 (Overall analysis).



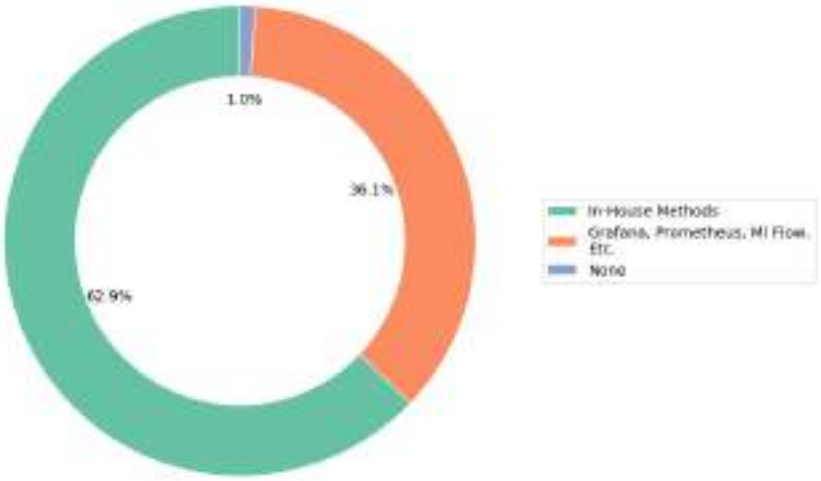

**Supplementary Figure 11.** Distribution plot of the responses from Q1.27 (Overall analysis).



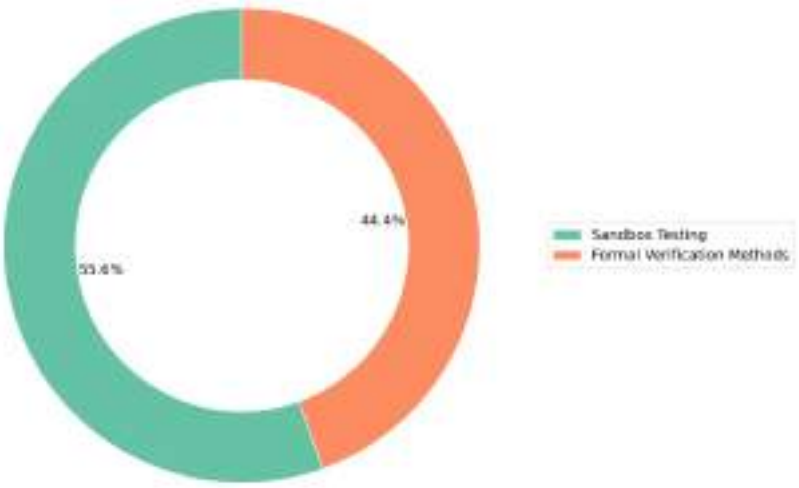

**Supplementary Figure 12.** Distribution plot of the responses from Q1.28 (Overall analysis).




**Supplementary Figure 13.** Distribution plot of the responses from Q1.29 (Overall analysis).



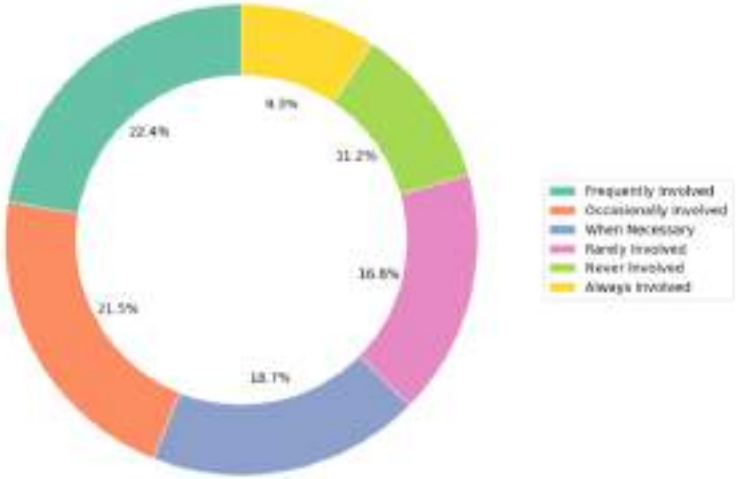

**Supplementary Figure 14.** Distribution plot of the responses from Q1.30 (Overall analysis).



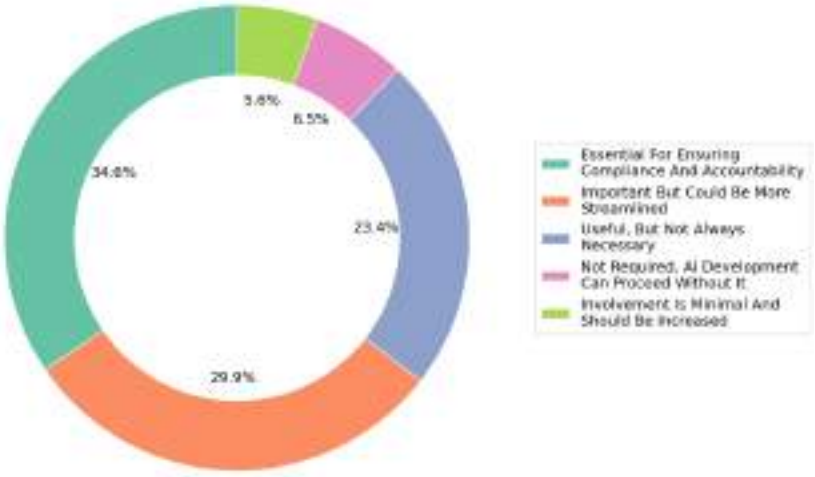

**Supplementary Figure 15.** Distribution plot of the responses from Q1.31 (Overall analysis).



**Supplementary Figure 16.** Distribution plot of the responses from Q2.4 (Overall analysis).

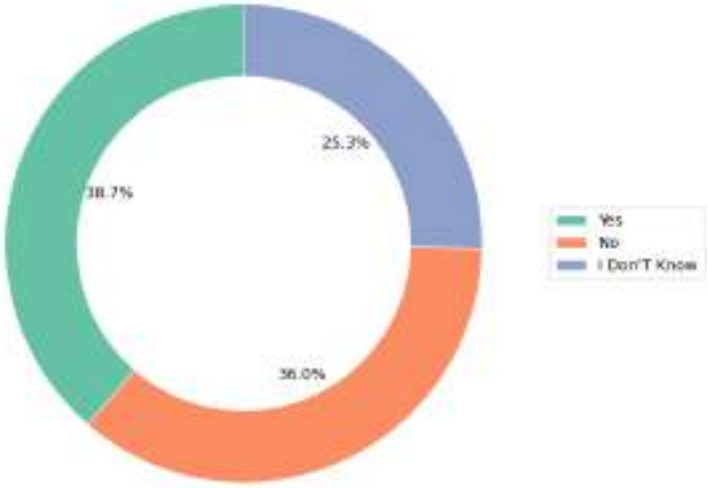



**Supplementary Figure 17.** Distribution plot of the responses from Q3.1 (Overall analysis).

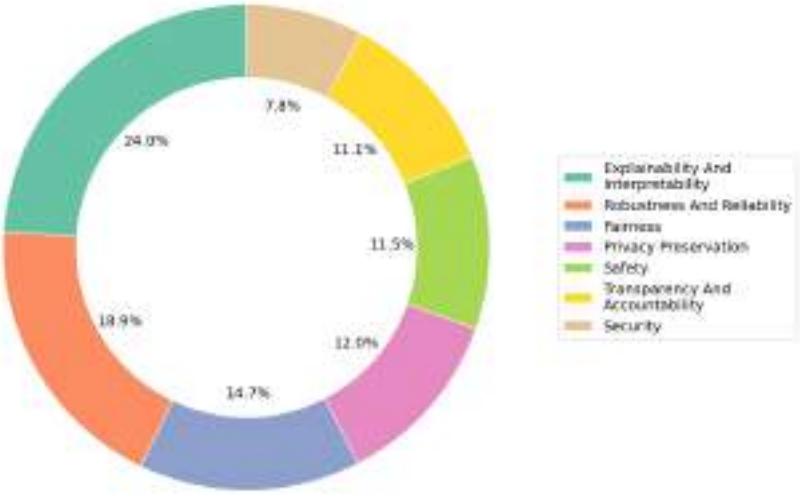



**Supplementary Figure 18.** Distribution plot of the responses from Q3.2 (Overall analysis).

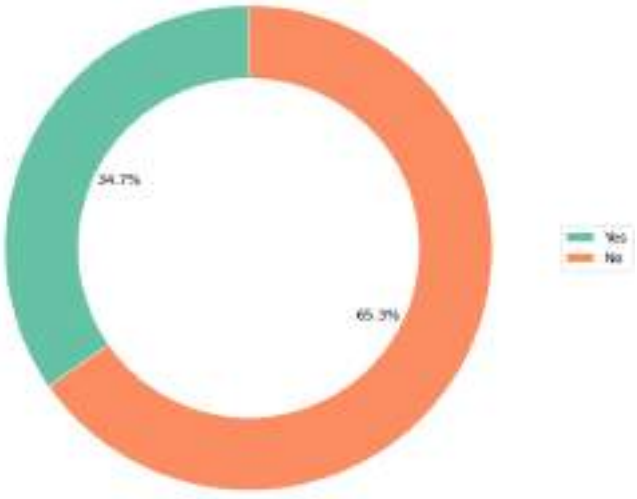
B2. Subgroup 1




**Supplementary Figure 19.** Distribution plot of the responses from Q1.32 (Subgroup 1 analysis).




**Supplementary Figure 20.** Distribution plot of the responses from Q2.1 (Subgroup 1 analysis).




**Supplementary Figure 21.** Distribution plot of the responses from Q2.2 (Subgroup 1 analysis).

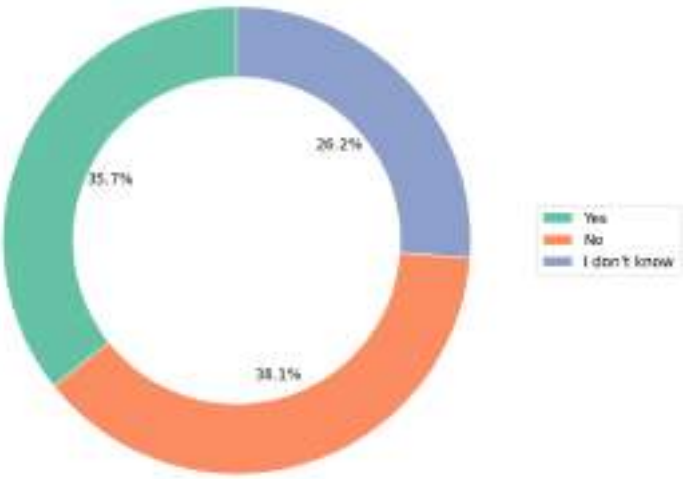



**Supplementary Figure 22.** Distribution plot of the responses from Q2.6 (Subgroup 1 analysis).

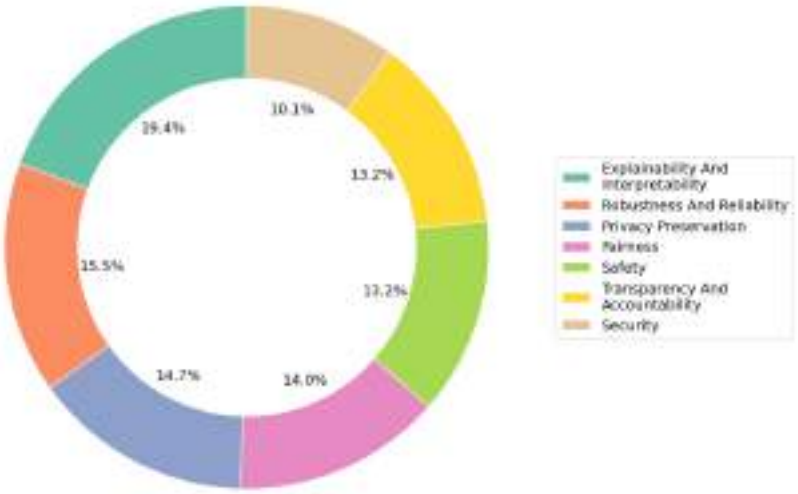



**Supplementary Figure 23.** Distribution plot of the responses from Q2.7 (Subgroup 1 analysis).

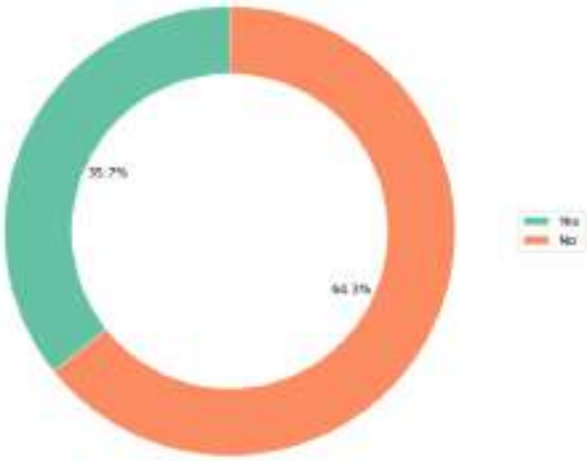
B3. Subgroup 2




**Supplementary Figure 24.** Distribution plot of the responses from Q1.32 (Subgroup 2 analysis).




**Supplementary Figure 25.** Distribution plot of the responses from Q2.1 (Subgroup 2 analysis).




**Supplementary Figure 26.** Distribution plot of the responses from Q2.2 (Subgroup 2 analysis).




**Supplementary Figure 27.** Distribution plot of the responses from Q2.6 (Subgroup 2 analysis).




**Supplementary Figure 28.** Distribution plot of the responses from Q2.7 (Subgroup 2 analysis).

## C. The completed KB enhancement templates per LSP

### C1. LSP1



**Supplementary Figure 29.** The completed KB enhancement template of LSP1.

### C2. LSP2



**Supplementary Figure 30.** The completed KB enhancement template of LSP2.

C3. LSP3



**Supplementary Figure 31.** The completed KB enhancement template of LSP3.

C4. LSP4



**Supplementary Figure 32.** The completed KB enhancement template of LSP4.

C5. LSP5



**Supplementary Figure 33.** The completed KB enhancement template of LSP5.

C6. LSP6



**Supplementary Figure 34.** The completed KB enhancement template of LSP6.

C7. LSP7



**Supplementary Figure 35.** The completed KB enhancement template of LSP7.