Ref. Ares(2026)9138516 - 28/09/2026



D3.4

Validation Report – v2

Related Work Package	**WP3 – Communication, Dissemination, Exploitation & Impact**
Related Task	Task 3.3
Lead Beneficiary	FORTH
Contributing Beneficiaries	All partners involved in WP3
Document version	v.1.0
Deliverable Type	R-Document, Report
Distribution level	PU-Public
Contractual Date of Delivery	30/09/2026
Actual Date of Delivery	28/09/2026

Contributors	Dimitrios I. Fotiadis (FORTH), Manolis Tsiknakis (FORTH), Kostantinos Marias (FORTH), Grigoris Kalliatakis (FORTH), Nikolaos S. Tachos (FORTH), Vasileios C. Pezoulas (FORTH), Eleni Tsalapati (ATC), G. Giotis (ATC), I. Lazaridis (ATC), M. Chatziathanasiou (ATC), Nineta Polemi (trustilio), Kitty Kioskli (trustilio), Theofanis Fotis (trustilio), George Douvitsas (trustilio), Panagiota Kontari (trustilio), Nic Fair (UOS), Paul Smart (UOS), Betul Gokya (UOS), Sam Senior (UOS), Steve Taylor (UOS)
Reviewers	KUL, UNIFI

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or Directorate-General for Communications Networks, Content and Technology. Neither the European Union nor the granting authority can be held responsible for them.

Version history

Version	Description	Date completed	Contributors
v0.1	TOC circulation	16/04/2026	UOS, FORTH
v0.2	Drafting of Section 1 introduction	06/08/2026	FORTH
v0.3	Drafting of Section 2 content	11/08/2026	UOS, FORTH
v0.4	Refinement of Section 2 content	18/08/2026	FORTH
v0.5	Further refinement of Section 2	21/08/2026	FORTH, ATC, Trustilio
v0.6	Drafting of Section 3 content	26/08/2026	ATC, CNR
v0.7	Further drafting of Section 3	02/09/2026	ATC
v0.8	Refinement of Section 3 content	08/09/2026	CNR, FORTH
v0.9	Drafting of Section 4 conclusions	11/09/2026	CNR
v1.0	Final version	28/09/2026	FORTH

Statement of Originality

This deliverable contains original unpublished work except where clearly indicated otherwise.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

Disclaimer

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

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

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

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

Executive summary

This deliverable, D3.4 Validation Report v2, is the second in a series of three Validation Reports produced under WP3 Task T3.3 (Evaluation of Methodology (NIST) and Policy Recommendations). It reports on the first full application of the validation plan defined in D3.3, covering both the FAITH Trustworthiness Assessment Framework version 2 (TAF v2) together with its supporting Knowledge Base and ontologies, and the suite of in isolation trustworthiness assessment tools used across the FAITH Large Scale Pilots (LSPs). The report first presents the development and validation of the FAITH AI Trustworthiness Ontology (AITO) and the TrustGuard Ontology, which together operationalise the extended FAITH Knowledge Base and enable the in context capabilities introduced in TAF v2. AITO was assessed through automated logical and structural validation, representational adequacy exercises, and ongoing domain coverage review, and was found to satisfy the formal requirements of the production pipeline while remaining subject to continued extension as new resources and use cases are incorporated. The TrustGuard Ontology was then validated through a set of stakeholder competency questions and an accompanying query tool, confirming that the ontology supports the practical information retrieval needs of the FAITH tool suite while also surfacing refinements that feed back into AITO itself. The report then presents the validation of TAF v2 as a method, conducted through two structured expert workshops involving developers, risk assessors, and end users across the FAITH LSPs. Both workshops converged on a consistent rating of validated with minor refinements, with participants confirming that the framework is conceptually clear, complete, and internally coherent. The refinements identified, including clearer phase to phase narrative, simplified terminology, and additional worked examples, are operational in nature rather than indicating any fundamental weakness in the framework. Validation of the framework's in context, whole system capabilities remains provisional, as the underlying System Security Modeller (SSM) tool was not yet available to participants at the time of the workshops. In parallel, the three in isolation trustworthiness assessment tools, namely the AI Model Hub and Passport, TrustSense, and TrustGuard, were validated through standardised formal feedback surveys administered across all seven LSPs, supplemented by informal feedback gathered through training sessions and partner exchanges. All three tools achieved a consistent expert assessment of validated with minor refinements, supported by strong scores for perceived usefulness, ease of use, and user trust. Three cross cutting priorities emerged across the tools and the framework alike: improving accessibility for non specialist users, reducing friction in integrating the tools into existing organisational workflows, and deepening cross tool integration through shared authentication infrastructure and alignment with the FAITH ontology. Taken together, the validation activities reported in this deliverable confirm that the FAITH methodology and toolset provide a credible, usable, and sufficiently useful basis for AI trustworthiness assessment in real world contexts. The identified refinements are already reflected in updates implemented or planned for Version 2 of the tools, and a further round of validation is planned once the SSM and updated TrustGuard are released, to directly assess the in context capabilities that could not be evaluated in this round.

Table of Contents

1. Introduction *9*
1.1 Objectives and Scope 9
1.2 Structure of the document 10
2. FAITH TRUSTWORTHINESS ASSESSMENT FRAMEWORK v2 AND ONTOLOGY DEVELOPMENT AND VALIDATION *11*
2.1 Introduction 11
2.2 FAITH AI Trustworthiness Ontology (AITO) - Design and Validation 11
2.2.1 FAITH AI Trustworthiness Ontology Design Decisions and Methodology 12
2.2.2 AITO Validation Activities 17
2.2.3 Validation Results 19
2.2.4 Ongoing Domain Coverage 24
2.2.5 Overall Findings and Implications 24
2.3 AITO Implementation in FAITH 24
2.3.1 Introduction 24
2.3.2 The TrustGuard Ontology Extension 25
2.3.3 TrustGuard Ontology Validation – Competency Questions 25
2.3.4 The TrustGuard Ontology Query Tool 27
2.3.5 TrustGuard Ontology Deployment Adaptations 31
2.3.6 Summary 32
2.4 FAITH AI TAF-v2 Validation 32
2.4.1 Development of AI TAF-v2 32
2.4.2 Methodology 33
2.4.3 TAF-v2 Validation Results 34
2.4.4 Key Actionable Findings 36
2.4.5 Updates 37
2.4.6 Summary of the validation outcomes 38
3. FAITH IN-ISOLATION TRUSTWORTHINESS ASSESSMENT TOOLS VALIDATION *40*
3.1 Methodology 40
3.2 AI Model Hub and Passport 41
3.2.1: Formal & Informal LSP Feedback Results 41
3.2.2 Analysis: Key Actionable Findings 47
3.2.3 Updates Implemented & Planned 48
3.3 TrustSense 50
3.3.1: Formal & Informal LSP Feedback Results 50
3.3.2 Analysis: Key Actionable Findings 52
3.3.3 Updates Implemented & Planned 52
3.4 TrustGuard 53
3.4.1 Formal & Informal LSP Feedback Results 53
3.4.2 Analysis: Key Actionable Findings 59
3.4.3 Updates Implemented & Planned 60
3.5 FAITH Tools Validation Summary 61
4. Conclusions and Next Steps *64*
4.1 Overall validation conclusions 64
4.2 Future actions 64
4.3 Link to subsequent deliverables and support to sustainability and exploitation activities 65
References *67*

List of Abbreviations

Abbreviation	Explanation
AAI	Authentication and Authorization Infrastructure
AI	Artificial Intelligence
AITO	FAITH AI Trustworthiness Ontology
BFO	Basic Formal Ontology
CCO	Common Core Ontologies
EU	European Union
HLEG	(EU) High-Level Expert Group on AI
IRI	Internationalized Resource Identifier
LSP	Large Scale Pilot / Large-Scale Pilot
OWL	Web Ontology Language
RMF	Risk Management Framework (as in NIST AI RMF)
SSM	System Security Modeller
TAF-v1	FAITH AI Trustworthiness Assessment Framework, version 1
TAF-v2	FAITH Trustworthiness Assessment Framework, version 2
TAM	Technology Acceptance Model
TWA	Trustworthiness Attribute

List of Tables

Table 1: Summary metrics for the AITO 19
Table 2: Taxonomic metrics for selected FAITH AITO branches 19
Table 3: Automated validation results for the AITO 20
Table 4: Vulnerabilities associated with the Oracle Threat 28
Table 5: Threats and Vulnerabilities associated with the Explainability trustworthiness attribute 30
Table 6: Comparison of complete merged and Lite TrustGuard ontology artifacts 31
Table 7 Summary of demographic and contextual characteristics of survey respondents (N=13, excluding one test entry). Counts marked with ~ are approximate, read from bar charts in the original survey report. 42
Table 8 Qualitative themes drawing from Q10/Q15/Q26. 48
Table 9 Perceived Usefulness Response Frequencies 50
Table 10 Interface and Workflow Ease of Use Response Frequencies 50
Table 11 Maturity level results 51
Table 12 Response Frequencies for the Human Factor dimension 51
Table 13 Summary of demographic and contextual characteristics of survey respondents for TrustGuard (N=13). Counts marked with ~ are approximate, read from bar charts in the original survey report 54

List of Figures

Figure 1: FAITH Validation Plan 9
Figure 2: AITO Architecture. Layered ontology architecture used within FAITH. 13
Figure 3: Examples of AITO Classes: Selected portions of the AI Model and AI Dataset class hierarchies following Taxonomy Walker-assisted taxonomic expansion. 15
Figure 4: Example AITO representation of the FLUX.2-klein-9B model and associated descriptive information, including its Hugging Face repository name and parameter count. 21
Figure 5: Example representation of the FLUX.2-klein-9B production process, including constituent processes and input/output relationships 22
Figure 6: Representation of a model-merging process in which Llama-2-70B and WizardLM-70B are inputs and Dolphin-2.5-70B is the resulting merged model 22
Figure 7: Example AITO representation of the ImageNet dataset, illustrating relationships between a dataset, constituent image content, and associated information-bearing artifacts 23
Figure 8: the TrustGuard Query Tool configured to retrieve vulnerabilities associated with an Oracle Threat 27
Figure 9: the TrustGuard Query Tool configured to identify which Threats and Vulnerabilities are associated with the Explainability Trustworthiness Attribute 29
Figure 10 Demographic profile of the survey respondents. Top row: distribution by LSP sector and organisation. Bottom row: primary role, AI lifecycle phase, and deployment environment. 42
Figure 11 Response distributions and mean scores for Perceived Usefulness items (Q6–Q9, Q23). Bars represent percentage of respondents per response category. 43
Figure 12 Response distributions and mean scores for Ease of Use items (Q11–Q14). 44
Figure 13 Respondent ratings for the Trust & Confidence dimension of the FAITH AI Model Hub. 44
Figure 14 Respondent ratings for the Attitude & Social Influence dimension of the FAITH AI Model Hub. Items assess overall satisfaction, willingness to recommend the platform to colleagues, and perceived organisational benefit. 45
Figure 15 Respondent ratings for the Compatibility & Adoption dimension of the FAITH AI Model Hub. Items address integration with existing workflows, fit within current toolchains, and intention to continue using the platform. 45
Figure 16 Response distributions and mean scores for AI Trustworthiness items (Q24–Q25). 46
Figure 17 Mean scores per TAM dimension. 47
Figure 18 Radar plot overview of TAM dimension scores. 48
Figure 19 trustSense Perceived Usefulness. 50
Figure 20 trustSense scores examples. 51
Figure 21 Demographic profile of the survey respondents for TrustGuard. Top row: distribution by LSP sector and organisation. Bottom row: primary role, AI lifecycle phase, and deployment environment. 54
Figure 22 Response distributions for Perceived Usefulness items (Q6–Q9) for TrustGuard. Bars represent percentage of respondents per response category. 55
Figure 23 Response distributions for Easy of Use items (Q11–Q14) for TrustGuard. Bars represent percentage of respondents per response category. 56
Figure 24 Response distributions for Easy of Use items Q17, Q18 for TrustGuard. Bars represent percentage of respondents per response category. 56
Figure 25 Response distributions for Easy of Use items (Q19–Q21) for TrustGuard. Bars represent percentage of respondents per response category. 57
Figure 26 Response distributions for Easy of Use items (Q22–Q24) for TrustGuard. Bars represent percentage of respondents per response category. 58

Figure 27 Response distributions for Easy of Use items (Q25–Q28) for TrustGuard. Bars represent percentage of respondents per response category. 59

1. Introduction

1.1 Objectives and Scope

This deliverable is the second in a series of three Validation Reports generated as part of the WP3 T3.3 (Evaluation of Methodology (NIST) & Policy Recommendations) activities.

The aim of this deliverable is to report on the first application of the validation plan reported in D3.3 – Validation Report, Chapter 3 and summarised in Figure 1: FAITH Validation Plan below. Firstly, in line with Tier 1, the validation activities associated with the updated and extended FAITH Trustworthiness Assessment Framework version 2 (TAF-v2) and the extended FAITH Knowledge Base will be detailed. In order to do this, it is also necessary to report on the development and validation of the domain-level FAITH AI Trustworthiness Ontology (AITO) and the application-level TrustGuard Ontology as these operationalise the Knowledge Base and enable the in-context advancespresent in TAF-v2 to be reflected in the FAITH tools.

Secondly, in line with Tiers 2 & 3 the validation of the in-isolation trustworthiness assessment tools (AI Model Hub and Passport, TrustSense, and TrustGuard) by the LSP User partners and risk assessors through standardised formal feedback will be presented. It should be noted that although the feedback mechanisms did include validation against the relevant NIST MEASURES (Tier 2), a deeper and more wide-ranging Tier 2 – NIST Measures validation has already been extensively reported in D3.3 Chapter 2. It is also worth noting that details of version one of the FAITH AI Trustworthiness Assessment Framework (TAF-v1) and the initial FAITH Knowledge Base validation were reported in D2.2.

Finally, this deliverable specifically does not report on the Tier 4 System-Under-Test evaluations, which is beyond the scope of this WP as it resides within the domain of each LSP work package itself.



Figure 1: FAITH Validation Plan

1.2 Structure of the document

Following this introduction, Section 2 addresses Tier 1 of the FAITH Validation Plan. It reports on the design and validation of the FAITH AI Trustworthiness Ontology (AITO), including its underlying methodology, validation strategy, and current coverage, followed by the implementation of AITO within FAITH through the TrustGuard Ontology, its competency questions, and its associated query tool. Section 2 then presents the validation of the FAITH Trustworthiness Assessment Framework version 2 (TAF v2), covering its development, the methodology used for its validation, the results and key actionable findings arising from the validation workshops, the resulting updates, and a summary of the overall validation outcome.

Section 3 addresses Tiers 2 and 3 of the FAITH Validation Plan through the validation of the three in isolation trustworthiness assessment tools used across the FAITH Large Scale Pilots (LSPs), namely the AI Model Hub and Passport, TrustSense, and TrustGuard. For each tool, the section presents the methodology used to gather formal and informal feedback, the resulting quantitative and qualitative findings, and the updates implemented or planned in response. Section 3 concludes with a summary comparing validation outcomes across all three tools and identifying priorities that apply consistently across the FAITH tool suite.

Section 4 concludes the deliverable. It presents the overall conclusions arising from the validation activities, sets out the future actions planned in response to the findings, and describes the relationship between this deliverable and subsequent reporting and exploitation activities within the project.

2. FAITH TRUSTWORTHINESS ASSESSMENT FRAMEWORK v2 AND ONTOLOGY DEVELOPMENT AND VALIDATION

2.1 Introduction

This section will report on the validation of the FAITH AI Trustworthiness Assessment Framework Version 2 (TAF-v2), the development of which was reported in depth in D2.2. It will also report on the development and validation of the FAITH AI Trustworthiness Ontology (AITO) and the TrustGuard Ontology, which includes the extended FAITH Knowledge Base, and which constitute significant iterative development and major outcomes within WP2 in response to feedback and findings from the first round of Pilots.

2.2 FAITH AI Trustworthiness Ontology (AITO) - Design and Validation

Artificial Intelligence (AI) systems are complex systems whose development and operation involve a diverse range of entities, including AI models, datasets, algorithms, software components, development and evaluation processes, governance resources, and human and organisational actors. Assessing the trustworthiness of such systems requires information about these entities to be represented in a structured and semantically coherent manner, particularly when information must be integrated across different stages of the AI lifecycle and accessed by software tools for trustworthiness assessment, risk management, information retrieval, and decision support.

The AI Trustworthiness Ontology (AITO) has been developed to provide a formal semantic framework for representing entities associated with AI systems and their trustworthiness.[1] Rather than functioning simply as a controlled vocabulary or taxonomy, AITO uses the Web Ontology Language (OWL) to provide formally specified classes, relations, and axioms, thereby supporting semantic interoperability, ontology-based reasoning, and semantic querying.

Within FAITH, the need for such a semantic framework arises from the project's objective of developing and validating a human-centric ecosystem for assessing, managing, and improving the trustworthiness of AI and robotics systems. FAITH seeks to assess trustworthiness systematically throughout the AI lifecycle rather than at only a single stage of system development. The project's AI Trustworthiness Assessment Framework (v2) is intended to support the identification of threats and vulnerabilities, estimation of trustworthiness-related risks, and selection of appropriate controls and mitigation measures across connected systems and activities including data preparation, model development, deployment, operation, and monitoring.

These requirements create a need to represent not only abstract trustworthiness concepts but also the entities that constitute AI systems and participate in their development and operation. Accordingly, the present report places particular emphasis on the following AITO modules and their associated classes and branches:

Module: AI Models, Systems and Algorithms

- **AI Model:** An Information Content Entity that specifies the computational structure and characteristic behaviour of an artificial intelligence system and that generically depends on one or more AI model artifacts.

[1] See https://git.soton.ac.uk/ontology/aito.

- **AI Dataset:** A Media Content Entity that is an organised collection of data items arranged according to a defined schema and curated for the training, validation, or evaluation of artificial intelligence models.
- **AI Capability:** A realizable entity that inheres in a material artefact designed to implement artificial intelligence methods.
- **AI Algorithm:** An Algorithm that prescribes a computational process that trains, fine-tunes, evaluates, deploys, operates, or maintains an AI model or AI system.

Module: Governance

- **AI Governance Document:** A Document Content Entity that sets out policies, procedures, controls, or requirements for the governance, oversight, regulation, compliance, or assurance of artificial intelligence models or systems.

Module: Stakeholders

- **AI Stakeholder Role:** A role that is borne by an entity in virtue of that entity having legitimate interests, rights, or liabilities that are affected or potentially affected by the design, development, deployment, operation, or decommissioning of an artificial intelligence system.

Module: AI Processes; Branch: Lifecycle Processes

- **AI Artifact Production Process:** An AI Artifact Lifecycle Process that creates, modifies, or prepares an AI Artifact.

Together, these classes and branches provide a structured representation of important AI artifacts, informational resources, capabilities, specifications, governance resources, stakeholder roles, and lifecycle processes.

The following sections of this report will describe how the FAITH-relevant portions of AITO were designed, developed, and validated. They present the principal ontology design decisions and development methodology, describe the logical, structural, and representational validation activities applied to the ontology, and summarise the results of this ongoing validation. A subsequent section describes the TrustGuard ontology extension and its role in supporting information requirements arising from the FAITH Trustworthiness Assessment Framework.

FAITH has benefitted from the full AI Trustworthiness Ontology (AITO) that has been developed in parallel with the Horizon Europe Themis 5.0 project as the work to develop the ontology was significant and directly useful to both projects. Each project maintained a clear separation of module and branch development, with the Themis project focusing on modules for Trustworthiness, Risks / Threats / Vulnerabilities, and additional AI Processes classes and branches, while FAITH focused on those modules and branches detailed above. In addition, the AITO has been differently deployed within the projects, with FAITH using the domain-level AITO to drive tool integration and data field harmonisation between the AI Model Hub, TrustGuard and SSM; to underpin the development of the application-level TrustGuard Ontology; and to enable the integration of the extended Knowledge Base for dynamic question selection within TrustGuard-v2. The complete ontology is available as a public resource under a permissive open access CC-BY license at https://git.soton.ac.uk/ontology/aito.

2.2.1 FAITH AI Trustworthiness Ontology Design Decisions and Methodology

2.2.1.1 Ontology Layers

AITO has been developed within a layered ontology architecture. At the upper level, it is aligned with the Basic Formal Ontology (BFO), which provides domain-independent categories for distinguishing fundamental kinds of entity (Arp et al., 2015; Otte et al., 2022). AITO also makes use of the Common Core Ontologies (CCO), which provide reusable mid-level concepts and relations that specialise the

more general categories introduced by BFO (Jensen et al., 2024). AITO builds on these foundations by introducing concepts specifically required for representing AI systems and AI trustworthiness. Where an appropriate concept or relation is already available through BFO or CCO, the preferred design strategy is to reuse that resource rather than introduce an AITO-specific equivalent.

The foundational distinctions provided by BFO help determine how AI-related concepts should be represented. One of the most important is the distinction between continuants, which persist through time while potentially undergoing change, and occurrents, which unfold or occur through time. This distinction enables, for example, AI models and datasets to be differentiated from the processes through which they are created, trained, evaluated, deployed, or otherwise manipulated. Similarly, an AI capability can be distinguished from a process in which that capability is realized. CCO provides more specialised resources concerning entities such as artifacts, agents, information entities, and processes, which can in turn be further specialised within AITO.



Figure 2: AITO Architecture. Layered ontology architecture used within FAITH.[2]

The resulting architecture distinguishes four levels of ontological representation, as illustrated in Figure 2. BFO provides the foundational layer, CCO supplies reusable mid-level semantics, and AITO provides the domain-level representation of AI and AI trustworthiness. AITO can, in turn, provide the semantic foundation for application-level ontologies developed to address more specialised operational requirements. Within FAITH, this application layer is represented by the TrustGuard ontology, which specialises AITO to support the FAITH Trustworthiness Assessment Framework. This

[2] Basic Formal Ontology (BFO) provides the foundational ontology, the Common Core Ontologies (CCO) provide reusable mid-level semantics, the AI Trustworthiness Ontology (AITO) provides the domain-level representation of AI and AI trustworthiness, and TrustGuard provides an application-level specialization supporting the FAITH Trustworthiness Assessment Framework.

layered approach is intended to promote conceptual consistency, semantic interoperability, modularity, and reuse while avoiding unnecessary duplication of concepts.

2.2.1.2 Analysis of AI Resources

Development of the FAITH-relevant portions of AITO (FAITH AITO) was informed by analysis of concrete AI resources rather than proceeding exclusively from an abstract characterisation of the AI domain. In particular, models and datasets hosted on the Hugging Face platform were examined as representative examples of the kinds of AI resources that the ontology should be capable of describing.

Hugging Face hosts a large and heterogeneous collection of AI models and datasets. Information about these resources is available through model and dataset cards and through structured metadata exposed by the platform's REST API. Model and dataset cards provide human-oriented descriptions covering characteristics such as intended purpose, technical features, development, limitations, and use, while API-accessible metadata provides structured information used to describe and categorize resources on the platform.

Analysis of these resources provided a basis for identifying distinctions relevant to the AITO representation of AI models and datasets. The objective was not to reproduce the Hugging Face metadata schema within AITO. Platform metadata is primarily intended to support the publication, discovery, documentation, and use of resources within a particular technical ecosystem, whereas AITO aims to provide a formally grounded representation of AI entities and their relationships. The Hugging Face analysis was therefore used to identify representational requirements and candidate distinctions that could be subjected to ontological analysis and, where appropriate, incorporated into the AITO class hierarchy.

This resource-oriented approach helps ensure that the relevant portions of AITO can characterise the kinds of models and datasets encountered in contemporary AI development while situating these resources within the broader ontological framework provided by AITO, CCO, and BFO.

2.2.1.3 Taxonomic Expansion

To support the systematic development of FAITH AITO class hierarchies, a custom Taxonomy Walker utility was developed to identify potential gaps in existing branches of the ontology. Given a target class, the utility traverses the class hierarchy and uses a suitably configured AI model (in this case, GPT-5.6) to identify candidate subclasses that may be missing.

For each class encountered during the traversal, the model is provided with information about the class and its existing direct subclasses and is prompted to suggest additional direct subclasses that could improve the completeness of the taxonomy. This process enables candidate gaps to be identified systematically at successive levels of the class hierarchy. Within FAITH the Taxonomy Walker was used to expand several branches concerned with AI resources, including AI Model, AI Dataset, AI Capability, and AI Governance Document. Figure 3 illustrates the resulting class hierarchies for AI Model and AI Dataset.

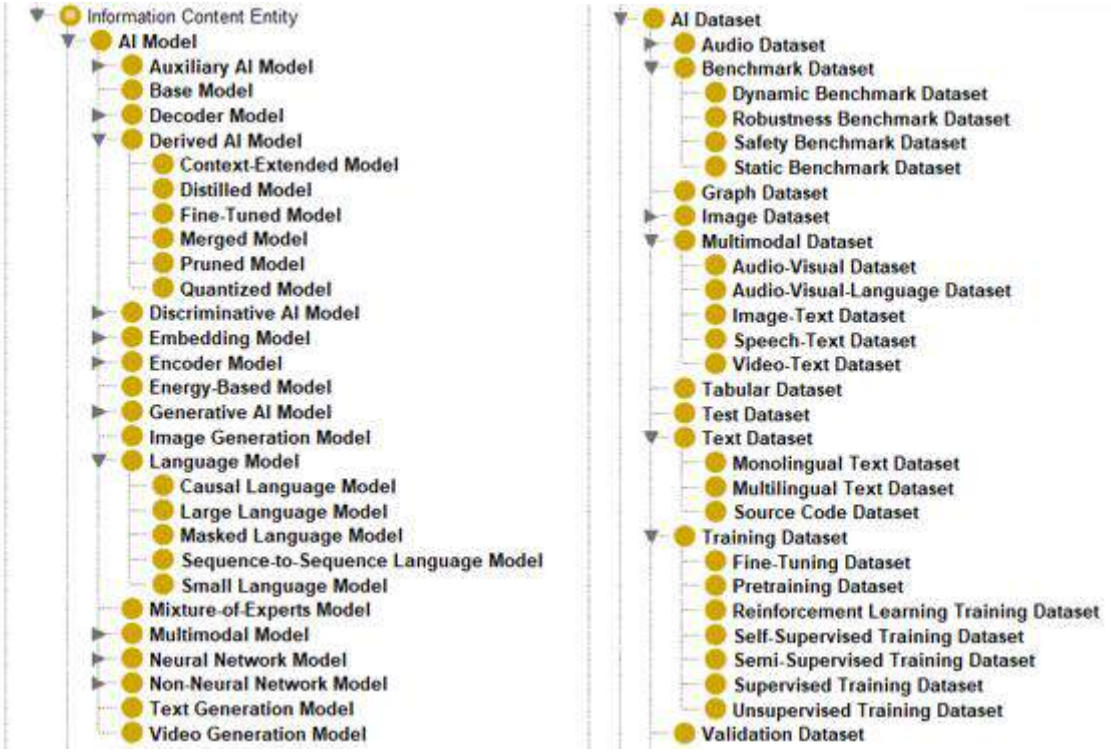


Figure 3: Examples of AITO Classes: Selected portions of the AI Model and AI Dataset class hierarchies following Taxonomy Walker-assisted taxonomic expansion.

The output of the Taxonomy Walker utility is not treated as authoritative ontology content. Instead, classes proposed by the model are reviewed by an ontology developer and may be rejected if, for example, they fall outside the intended scope of AITO, duplicate an existing concept, reflect a lexical rather than ontological distinction, occur at an inappropriate level of granularity, or cannot be situated coherently within the existing ontology architecture. Human review therefore remains an integral part of the development process.

The Taxonomy Walker utility therefore provides an AI-assisted mechanism for exploring potential gaps in an existing taxonomy rather than an automated mechanism for determining ontology content. It combines the ability of generative AI to identify candidate concepts across an extended class hierarchy with human responsibility for the ontological commitments ultimately incorporated into the ontology.

2.2.1.4 Class Definitions and Language Translation

A further requirement of FAITH AITO development concerns the provision of human-readable definitions for ontology classes. Definitions are important for ontology maintainability and for communicating the intended meanings of classes to users who need to interpret or extend the ontology. They also help distinguish classes whose labels may appear similar and make explicit the conceptual distinctions underlying the class hierarchy.

The ontology uses an AI-assisted approach to definition development, with GPT-5.6 used to generate candidate skos:definition annotations. Where appropriate, definitions follow an Aristotelian pattern based on genus and differentia: the definition identifies the broader category to which an entity belongs and then specifies the characteristics that distinguish it from other entities belonging to that category. This approach aligns naturally with the taxonomic organization of AITO, since the asserted superclass can provide the genus for the definition of a more specialised class.

As with taxonomic expansion, AI-generated definitions are treated as candidate ontology content rather than as automatically authoritative. Human review remains important for ensuring that a

proposed definition captures the intended meaning of the class, is compatible with its position in the ontology, and does not introduce commitments that are absent from or inconsistent with the formal representation. Human-readable class names and definitions are represented systematically using rdfs:label and skos:definition, respectively.

To improve the accessibility of the ontology within a multilingual European context, the documentary content of the ontology has also been translated into a range of European languages. Google Cloud Translation was used to generate translations of the English-language rdfs:label and skos:definition annotations. The resulting annotations use appropriate language tags, enabling ontology software and semantic applications to distinguish and select labels and definitions according to language.

Rather than incorporating all translated annotations directly into the core AITO ontology, the multilingual annotations are maintained in a separate ontology, AITO_Languages.owl. This keeps the core ontology comparatively compact while allowing multilingual information to be incorporated when required. The language ontology can therefore be used alongside AITO by applications that require localised presentation of ontology terminology, while applications requiring only the core semantic representation need not process the additional annotations.

The use of AI-assisted definition generation and machine translation provides a scalable mechanism for maintaining the documentary and multilingual layers of AITO as the ontology develops. Both activities concern the human-readable annotation layer rather than altering the formal OWL axioms that determine the logical semantics of the ontology.

2.2.1.5 Programmatic Ontology Construction

The FAITH AITO was produced using a programmatic ontology-engineering workflow rather than being maintained primarily through direct manual editing of OWL files. Ontology content is developed using editable and intermediate resources and subsequently transformed into OWL artifacts through an automated production pipeline. This approach separates the resources used during ontology development from the formal ontology artifacts produced for publication, reasoning, and semantic querying.

The production workflow combines structured resources, including spreadsheets and JSON files, with a collection of software tools. Python scripts and Jupyter notebooks are used for activities such as data processing and transformation, while custom Java utilities built using the OWLAPI provide programmatic access to OWL ontology structures. Additional ontology-engineering tools, including ROBOT, are incorporated into the build and validation pipeline. The source resources and associated software are maintained under version control, providing a traceable basis from which AITO releases can be generated.

Programmatic construction was adopted to improve the consistency, maintainability, and reproducibility of ontology development. IRIs and annotations can be generated according to common conventions, changes to structured source resources can be propagated systematically into the resulting ontology, and large-scale modifications can be performed without repeated manual editing of OWL files. The approach also enables ontology construction to be integrated directly with automated quality-assurance procedures, so that generated artifacts can be subjected to structural and logical validation as part of the production process.

Overall, the workflow provides a reproducible path from editable development resources to validated OWL artifacts. As AITO evolves in response to new requirements and validation findings, updated releases can be regenerated using the same source-controlled resources, software, and production procedures.

2.2.2 AITO Validation Activities

Having been designed and constructed, it was then necessary for the FAITH AITO to be validated. This consisted of a structured strategic process commonly used for formal ontology validation, which included Logical Validation for internal consistency; Syntactic and Structural Validation for convention consistency; Representational Adequacy for sufficient domain characterisation; and Competency Questioning for the usability and usefulness of outputs. The first three stages will be discussed in this section, and the Competency Question validation work reported in TrustGuard Ontology Extension section.

2.2.2.1 Validation Strategy

Ontology validation was an integral part of the AITO development process and addressed multiple dimensions of ontology quality. An ontology may be syntactically well formed while containing logical inconsistencies or may be logically consistent while failing to represent important aspects of its intended domain. Validation must therefore consider both the formal properties of the ontology and its representational adequacy.

The AITO validation strategy combined automated and human-led activities. Automated procedures were used to assess properties such as logical consistency and conformity with structural requirements, while aspects requiring conceptual judgement were evaluated through human analysis. Representational adequacy, for example, can be assessed by attempting to use AITO to characterise concrete AI resources or scenarios and identifying any missing concepts or relationships.

Validation was integrated into an iterative development process rather than being performed only on a completed ontology. New or revised ontology content was validated, identified problems or gaps informed subsequent refinement, and the resulting ontology was revalidated.

2.2.2.2 Logical Validation

Logical consistency is a fundamental requirement for an OWL ontology intended to support automated reasoning. Ontology axioms can interact in ways that introduce contradictions or cause classes to become unsatisfiable, making logical validation an important part of the ontology-development process.

FAITH AITO incorporates logical validation within its automated production pipeline. The ROBOT ontology-engineering tool (Jackson et al., 2019) was used to process the generated ontology, with the HermiT reasoner used to perform logical consistency and class-satisfiability checking before a new version of AITO was treated as a valid ontology artifact.

The validation process also determined the OWL profile of the generated ontology. OWL profiles characterise the language constructs used by an ontology and provide information about its expressivity and associated reasoning requirements. In addition, reasoning was used to assess whether the ontology is logically consistent and whether its named classes remain satisfiable. An unsatisfiable class is one whose axioms or inherited restrictions prevent it from possessing instances without generating a contradiction.

Automating these checks provided a safeguard against logical problems that may not be apparent through manual inspection, particularly as the FAITH AITO grows and axioms interact across different parts of the ontology and its imported resources. Logical consistency nevertheless represents only one dimension of ontology quality: a consistent ontology may still contain inappropriate classifications, missing concepts, or inadequate modelling structures.

2.2.2.3 Syntactic and Structural Validation

Syntactic and structural validation concerns whether ontology entities have been generated and documented according to the conventions adopted for AITO. The programmatic ontology-

development methodology described above enables many of these requirements to be checked systematically as part of the ontology-production pipeline.

One basic requirement concerns identifiers. Each ontology entity must possess a unique Internationalized Resource Identifier (IRI) so that it can be referenced unambiguously within AITO and by external semantic applications. Programmatic generation enables IRIs to be assigned according to consistent conventions and duplicate identifiers to be avoided or detected.

FAITH AITO entities are also required to possess human-readable annotations. In particular, class names are represented using rdfs:label, while definitions are represented using skos:definition. The production pipeline provides a means of checking that these annotations are present and applied consistently. Structural validation can additionally include checks concerning entity declarations, namespace usage, and successful serialization of the ontology as RDF/OWL.

Integrating these checks into the ontology-production process enabled structural problems to be identified before a generated artifact was released. The approach also allowed further automated checks, such as the detection of duplicate labels, malformed identifiers, or deprecated entities, to be incorporated as validation requirements evolved.

2.2.2.4 Representational Adequacy

Logical and structural validation do not establish that an ontology contains the conceptual resources required for its intended applications. The FAITH AITO may have been validated as logically consistent and structurally well formed but may nevertheless lack a class or relation needed to describe a particular AI system. A further validation criterion is therefore representational adequacy: the extent to which the classes, relations, and modelling structures provided by an ontology enable relevant entities and states of affairs to be represented.

In the context of FAITH, the central question is whether the AITO can adequately characterise relevant AI artifacts, informational resources, capabilities, and production processes. This can be evaluated through concrete modelling exercises in which a representative resource or scenario is described using AITO. Such exercises can identify information that can be represented using existing ontology resources as well as information for which an appropriate class, relation, or modelling structure is absent.

A publicly available AI model provided one useful test case. In the validation activities reported in Section 2.2.3.4, the FLUX.2-klein-9B model is used to examine whether AITO can represent information concerning a concrete AI model and its production process. Comparable exercises can be undertaken for datasets, algorithms, AI capabilities, governance documents, and other AI production processes.

Representational testing can reveal gaps that are difficult to identify by examining the class hierarchy alone. AITO may, for example, contain appropriate classes for a model and dataset but lack a relation required to connect them in the intended manner. This approach complements the analysis of Hugging Face resources described in Section 2.2.1.2: whereas that analysis helped identify the distinctions required to characterise contemporary AI resources, representational-adequacy testing evaluated whether the resulting ontology could successfully represent concrete instances of those resources.

Representational adequacy is necessarily an ongoing validation activity, since no finite set of examples can establish adequacy for every AI resource or application. An expanding collection of representative cases can instead be used to test different portions of AITO and identify requirements for further refinement. Over time, these cases could form a reusable representational-adequacy test suite against which successive versions of AITO can be evaluated.

2.2.3 Validation Results

2.2.3.1 Introduction

Validation of AITO is an ongoing process rather than an activity performed only after ontology development has been completed. AITO continues to evolve as its conceptual coverage is extended, modelling patterns are refined, and new requirements emerge from the projects and applications that use the ontology. The results presented in this section therefore characterise the current state of AITO rather than constituting a final certification of ontology quality.

Automated validation procedures provide baseline quality assurance for generated ontology artifacts, while complementary activities, including representational-adequacy testing and domain-coverage analysis, help identify areas requiring further development. Validation results consequently serve two purposes: establishing that the current ontology satisfies defined formal and structural requirements, and identifying opportunities for subsequent refinement as FAITH, AITO and the AI domain continue to evolve.

2.2.3.2 Current Ontology Metrics

Ontology metrics provide a quantitative characterisation of the scale and structure of AITO. Although such metrics do not by themselves establish conceptual correctness or representational adequacy, they provide a useful description of the ontology artifact and a baseline against which subsequent versions can be compared. The AITO release associated with the present deliverable contains 23,277 axioms, including axioms contained in its imported ontologies, of which 4,890 are logical axioms.[3] The ontology includes 1,277 classes, 24 object properties, 108 individuals, and the additional entities summarized in Table 1.

Metric	Value
Total axioms	23,277
Logical axioms	4,890
Classes	1,277
Object properties	24
Data properties	0
Annotation properties	1
Individuals	108

Table 1: Summary metrics for the AITO

Global metrics can be complemented by examining the portions of the class hierarchy that are particularly relevant to FAITH. Table 2 reports the number of direct subclasses and total descendants for seven of the principal FAITH-relevant classes.

FAITH AITO branch	Direct subclasses	Total descendants
AI Model	17	74
AI Dataset	11	42
AI Capability	9	65
AI Algorithm	8	58
AI Governance Document	18	79
AI Stakeholder Role	14	176
AI Artifact Production Process	3	24

Table 2: Taxonomic metrics for selected FAITH AITO branches

[3] These numbers include the axioms contained in imported ontologies.

These figures demonstrate that the FAITH AITO extends beyond a small collection of high-level AI concepts and contain substantial taxonomic differentiation. AI Stakeholder Role has the largest number of descendants among the branches reported here (176), followed by AI Governance Document (79), AI Model (74), AI Capability (65), and AI Algorithm (58). The figures also illustrate the difference between immediate taxonomic structure and the deeper hierarchy: for example, AI Stakeholder Role has 14 direct subclasses but 176 descendants overall.

These metrics should not, however, be interpreted as direct measures of ontology quality or domain coverage. The number of subclasses required for an appropriate representation depends upon the structure and granularity of the domain being modelled. Their primary purpose is therefore to document the current state of the ontology and to provide quantitative context for the validation activities reported here.

2.2.3.3 Automated Validation Results

Automated validation formed part of the AITO production pipeline and functions as a release gate: a new version of the ontology was generated for release only if it satisfied the required structural and logical validation checks. The availability of the AITO release associated with this deliverable therefore indicated that all mandatory automated validation criteria were satisfied.

Logical consistency and class satisfiability were evaluated using the HermiT reasoner through the ROBOT ontology-engineering tool. The released ontology was found to be logically consistent, with no unsatisfiable named classes. Profile analysis classified the ontology as OWL 2 DL. Structural checks also confirmed compliance with the requirements applied during ontology generation, including unique entity identifiers and the required human-readable annotations. The results are summarized in Table 3.

Validation criterion	**Result**
Successful ontology generation	PASS
OWL profile	OWL 2 DL
Logical consistency (HermiT)	PASS
Unsatisfiable named classes	0
Unique IRI requirements	PASS
Required rdfs:label annotations	PASS
Required skos:definition annotations	PASS

Table 3: Automated validation results for the AITO

These results establish that the released AITO artifact satisfies the formal and structural requirements enforced by the production pipeline. They do not, however, establish the completeness or representational adequacy of the ontology, which require the complementary forms of validation discussed elsewhere in this report.

2.2.3.4 Representational Adequacy

Representational adequacy concerns the extent to which AITO provides the classes, relations, and modelling structures required to represent concrete entities and states of affairs within its intended domain. To provide initial evidence concerning this criterion, a series of modelling exercises was undertaken using examples relating to AI models, model-production processes, and datasets. These exercises were not intended to establish that AITO can represent every relevant aspect of every AI system. Rather, they tested whether selected real-world states of affairs can be expressed using the current ontology and provided a means of identifying potential representational gaps.

The first exercise concerned the FLUX.2-klein-9B model and illustrated how AITO can represent an AI model together with descriptive information about that model. As shown in Figure 4, FLUX.2-klein-9B is represented as an instance of AI Model and is associated with information concerning its Hugging Face repository name and parameter count. These descriptive entities are themselves classified within the ontology: the repository name is represented using the AI Model Name hierarchy, while the parameter count is represented as an AI Model Parameter Count. The example therefore illustrates how a concrete AI model can be connected to informational entities that designate or describe it.

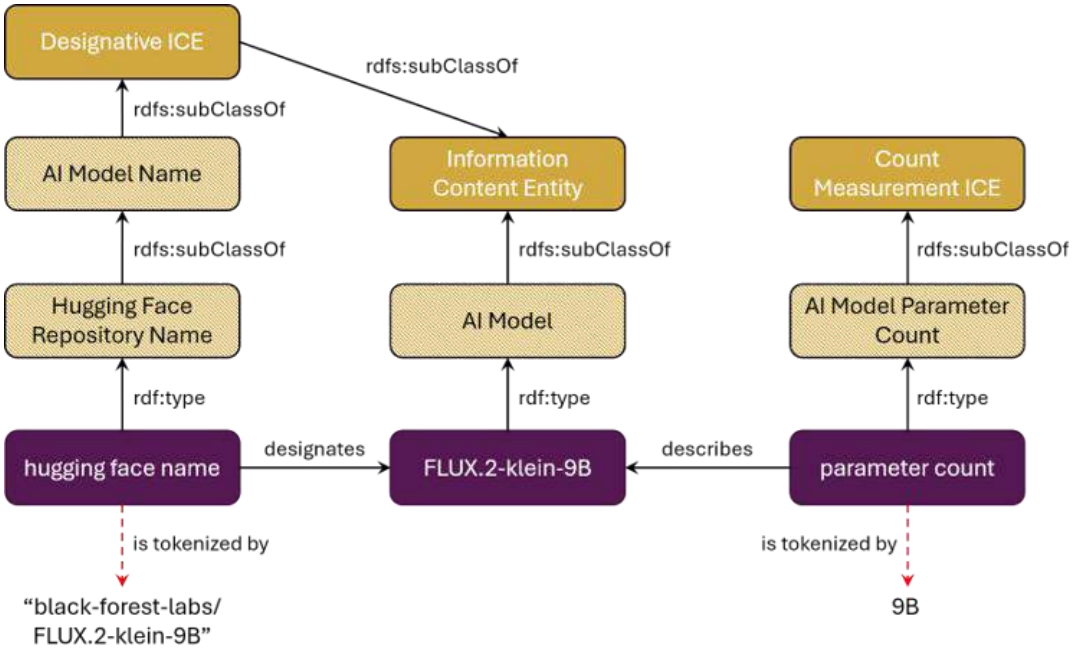


Figure 4: Example AITO representation of the FLUX.2-klein-9B model and associated descriptive information, including its Hugging Face repository name and parameter count.[4]

A second exercise extended the representation from the model itself to the process through which it is produced. Figure 5 represents the FLUX.2-klein-9B production process as an instance of AI Model Production Process comprising a sequence of constituent processes, including model training, model distillation, safety fine-tuning, external third-party evaluation, and final assessment. The representation used process-part relations to connect these activities to the overall production process and precedence relations to express their ordering. It also represented the training dataset as an input to the production process and FLUX.2-klein-9B as its output. This example demonstrates that AITO can represent not only AI artifacts but also structured lifecycle processes and the relationships between processes, their inputs, and their outputs.

[4] [Notation: solid gold nodes represent BFO/CCO classes; striped gold nodes represent AITO classes; purple nodes represent OWL individuals; black arrows represent standard relations; and blue arrows represent modal relations.]

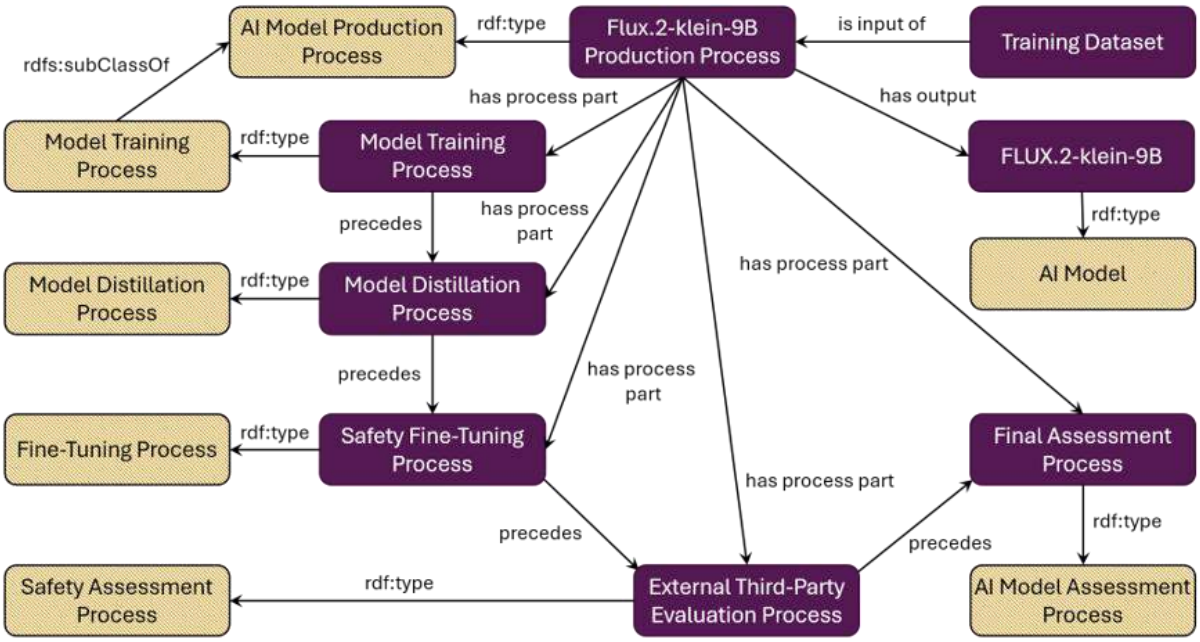


Figure 5: Example representation of the FLUX.2-klein-9B production process, including constituent processes and input/output relationships

The third exercise concerned model merging and tested the representation of a concrete model-transformation scenario. As shown in Figure 6, Llama-2-70B and WizardLM-70B were represented as inputs to a Model Merging Process whose output is Dolphin-2.5-70B. The example also illustrated the interaction between asserted information and the formal semantics of AITO. The Model Merging Process is defined as a form of AI Model Adaptation Process that takes two or more AI models as inputs and produces a merged model as output. The ontology additionally characterises a Merged Model in terms of being an AI model that is the output of a Model Merging Process. The representation therefore illustrates how formal axioms can support classification of entities on the basis of their participation in represented processes, rather than requiring every classification to be asserted independently.

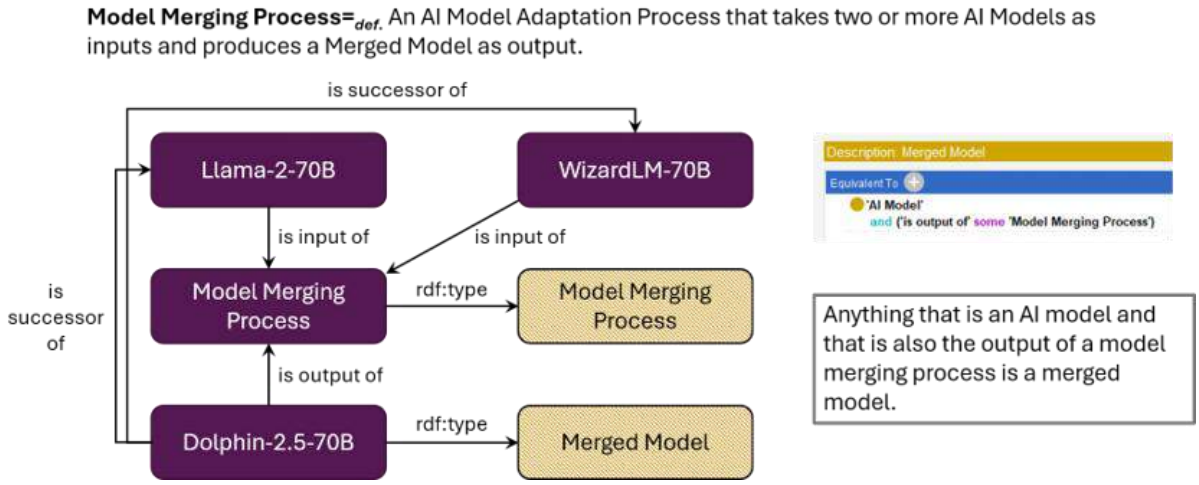


Figure 6: Representation of a model-merging process in which Llama-2-70B and WizardLM-70B are inputs and Dolphin-2.5-70B is the resulting merged model

The final exercise considered the representation of datasets. Figure 7 uses ImageNet to illustrate the distinction between an AI dataset as an informational entity and the artifacts on which its informational content generically depends. ImageNet Dataset is represented as an instance of Image

Dataset, itself a subclass of AI Dataset, while an individual image is represented as a constituent part of the dataset. The example further distinguishes the informational content of an image from a corresponding image file and the information-bearing artifacts on which informational content generically depends. This demonstrated the value of the BFO/CCO-aligned architecture in representing distinctions between informational entities, their constituent content, and their material realisations.

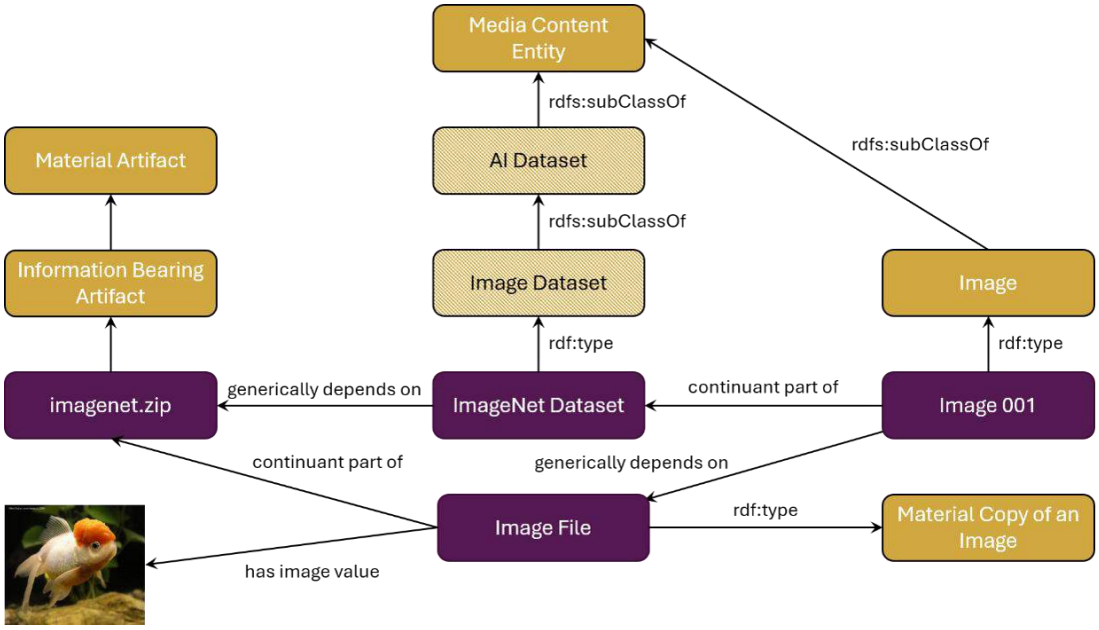


Figure 7: Example AITO representation of the ImageNet dataset, illustrating relationships between a dataset, constituent image content, and associated information-bearing artifacts

Taken together, these exercises provide initial evidence that AITO can represent several different kinds of state of affairs relevant to FAITH: the characteristics of an individual AI model; the processes involved in producing a model; transformations involving multiple models; and the structure and realisation of AI datasets. They also demonstrate that representational adequacy depends on more than the presence of appropriate classes in the taxonomy. Relations are required to connect entities into coherent representations, while formal axioms can enable additional information to be inferred from those representations.

These examples should nevertheless be regarded as initial representational tests rather than evidence of comprehensive domain coverage. Representational adequacy is necessarily an ongoing validation activity, since no finite collection of examples can demonstrate that AITO can represent every relevant AI resource or scenario. Future validation can extend the approach to additional models, datasets, governance resources, lifecycle processes, and application-specific scenarios. Cases derived directly from FAITH stakeholder and application requirements would be particularly valuable because they would test the ontology against concrete project information needs.

Over time, these modelling exercises could be developed into a reusable representational-adequacy test suite. Each case could specify independently motivated representational requirements and document the AITO classes and relations used to satisfy them. Cases that expose information that cannot be represented satisfactorily would identify candidate ontology gaps, while previously successful cases could be revisited following ontology revisions to ensure that established modelling capabilities have not been inadvertently lost. Representational testing can thereby form part of the continuing cycle through which application requirements inform the refinement and validation of AITO.

2.2.4 Ongoing Domain Coverage

Domain coverage concerns whether AITO contains a sufficiently comprehensive set of concepts for representing its intended domain. It is distinct from representational adequacy: whereas the modelling exercises described above test whether particular states-of-affairs can be represented using the ontology, domain-coverage analysis considers the breadth of concepts represented across the domain as a whole.

For the FAITH AITO, analysis of Hugging Face resources provides one mechanism for evaluating coverage. The diversity of models and datasets available through the platform exposes the ontology to a range of resource types and descriptive characteristics. Where a conceptually relevant distinction encountered in these resources cannot be mapped to an existing FAITH AITO concept, this provides evidence of a potential gap requiring further analysis.

Coverage cannot be assessed through lexical matching alone, however. Different terms may denote the same concept, while superficially similar terms may have different meanings. Evaluation therefore needs to consider definitions, superclass relationships, and other ontological context when determining whether an externally identified concept is already represented by AITO.

Continuing domain coverage analysis within FAITH will focus particularly on AI models, datasets, algorithms, capabilities, governance resources, stakeholder roles, and production processes. As additional AI resources and development practices are examined, identified concepts can be compared with FAITH AITO and used to inform further ontology development where appropriate.

2.2.5 Overall Findings and Implications

The validation results provide complementary evidence concerning different aspects of AITO quality. Automated validation establishes that the current ontology release satisfies the formal and structural requirements enforced by the production pipeline: the ontology conforms to OWL 2 DL, is logically consistent under HermiT reasoning, and contains no unsatisfiable named classes. The representational exercises provide complementary evidence that AITO can be used to characterise concrete AI resources and states of affairs, including individual models, model-production and adaptation processes, and datasets. Domain-coverage validation provides an additional mechanism for identifying areas in which the ontology may require further development.

These results should not be interpreted as demonstrating that AITO is complete. Logical and structural validation provide well-defined release criteria, whereas representational adequacy and domain coverage are necessarily progressive forms of validation that can be strengthened as additional cases and requirements are considered. The following section moves from validation of the domain ontology to its specific application within FAITH, describing how the TrustGuard ontology extends AITO to represent the FAITH Knowledge Base and supports the competency questions and information-retrieval requirements associated with the Trustworthiness Assessment Framework.

2.3 AITO Implementation in FAITH

2.3.1 Introduction

The FAITH AITO has been used to develop the TrustGuard Ontology. This is an application-level ontology developed to enable cross-tool integration, respond to initial feedback from the first round of Pilot Use Cases, and support the FAITH Trustworthiness Assessment Framework version2 (TAF-v2). It is deployed across the FAITH project to harmonise data fields in the updated AI Model Hub and Data Passport, version 2 of the TrustGuard tool, and the System Security Modeller (SSM – aka Spyderisk). This semantic and conceptual harmonisation enables closer and more efficient integration between these tools, where information can flow seamlessly between all three. In addition, the TrustGuard Ontology integrates the extended FAITH Knowledge Base. This supports the dynamic selection of

questions within the TrustGuard tool in order that every question presented to the user is directly relevant to the context in which the AI Model and/or Dataset is situated. This was in response both to the feedback from risk assessors during the Pilot 1 deployment and to align with the in-context aspects of the FAITH AI Trustworthiness Assessment Framework version2 (TAF-v2).

2.3.2 The TrustGuard Ontology Extension

The TrustGuard Ontology extends AITO by representing application-specific information concerning threats, vulnerabilities, controls, trustworthiness properties, asset types, lifecycle stages, and TrustGuard assessment questions as represented in the FAITH Knowledge Base. Whereas AITO provides reusable domain-level concepts concerning AI and AI trustworthiness, the TrustGuard Ontology specialises this semantic framework for the particular information requirements of FAITH and the TrustGuard tool.

The current scope of the TrustGuard Ontology is principally determined by the FAITH Knowledge Base containing structured information used by the TrustGuard tool in particular. The Knowledge Base records not only individual entities but also relationships between them. For example, vulnerabilities are associated with threats and lifecycle stages, controls can address vulnerabilities and apply to particular asset types, and trustworthiness properties and assessment questions are connected to other elements of the assessment framework. These relationships provide the basis for many of the information-retrieval requirements addressed by TrustGuard.

The Knowledge Base was transformed programmatically into the TrustGuard OWL Ontology. During this process, the Knowledge Base content was normalised, recurring entities were assigned consistent identifiers, and relationships encoded within the tabular structure are converted into RDF/OWL assertions. Concepts already represented within AITO were reused where appropriate, while application-specific resources were represented within the TrustGuard layer.

This approach allowed the Knowledge Base to remain an editable resource for project experts while providing an OWL representation suitable for semantic querying and reasoning. Programmatic generation also improved consistency, traceability, and reproducibility: when the Knowledge Base is updated, the TrustGuard Ontology can be regenerated using the same transformation rules rather than requiring a parallel ontology representation to be maintained manually.

The relationship between the Knowledge Base and TrustGuard is also important for validation. Because the ontology is generated systematically from the Knowledge Base, the relationships encoded in the Knowledge Base can be exposed through semantic queries over the TrustGuard Ontology. This does not by itself establish the correctness of the source information or of the ontological modelling decisions used to represent it. It does, however, provide a reproducible mapping between the source resource and its semantic representation, enabling the TrustGuard Ontology to function as an ontology-based interface for the information captured by the Knowledge Base.

2.3.3 TrustGuard Ontology Validation – Competency Questions

Competency questions provide an application-orientated mechanism for evaluating whether an ontology contains the information and semantic structures required to support its intended use. A competency question expresses an information requirement in natural language without initially prescribing how that information should be represented or retrieved. The question can subsequently be operationalised as a formal query against the ontology.

A set of 13 competency questions for the TrustGuard Ontology were co-created by various FAITH project stakeholders, including developers and risk assessors. These questions provide a particularly valuable basis for validation because they originate from the information needs of the project rather than being constructed solely from the existing ontology structure. The questions can be grouped into several related families, as follows.

2.3.3.1 Threats

A first group concerns the circumstances under which threats should be considered applicable:

- Given a domain (LSP), asset type, model architecture, model learning algorithm, model function, and lifecycle stage, which threats are applicable?
- Given a threat, which asset types make its vulnerabilities and/or controls applicable?
- What is the definition / description of a given threat?

These questions require the TrustGuard Ontology to represent not merely individual threats but the contextual relationships that determine their applicability and definition. In particular, they require relationships between threats, vulnerabilities, controls, asset types, lifecycle stages, and other characteristics of the AI resource under consideration.

2.3.3.2 Vulnerabilities

The second group concerns the retrieval and contextualization of vulnerabilities:

- Given a threat and a domain (LSP), model architecture, model learning algorithm, dataset type, and lifecycle stage, which vulnerabilities are applicable?
- Given a threat, which vulnerabilities are associated with it?

These questions exercise the many-to-many relationships encoded within the Knowledge Base. An ontology-based representation is well suited to such cases because an entity can participate in multiple explicitly represented relationships without the duplication associated with a row-oriented representation.

2.3.3.3 Trustworthiness

The next group concerns relationships between vulnerabilities, threats, and trustworthiness properties:

- Given a trustworthiness property, which threats and vulnerabilities are associated with it?
- Given a vulnerability, which trustworthiness properties are associated with it?
- What is the definition / description of a given trustworthiness property?

These questions are particularly important for the objectives of FAITH because they connect threat and vulnerability information with the trustworthiness characteristics that form the basis of the FAITH TAF-v2 and the TrustGuard tool.

2.3.3.4 Controls and Mitigations

Several competency questions concern controls and their relationship to vulnerabilities and threats:

- Given a vulnerability, which controls address/mitigate it?
- Given a vulnerability and an asset type, which controls are applicable?
- Given an asset type, which threats can be mitigated by the applicable controls?
- Which assessment question is associated with a given control?
- What is the definition / description of a given control?

Answering these questions requires relationships to be traversed across several kinds of entity. A control may address a vulnerability, the vulnerability may be associated with a threat, and the applicability of the control may depend upon the type of asset involved. The question concerning which threats a control is ultimately mitigating is therefore a good example of an information requirement that may require traversal of an indirect semantic path rather than retrieval of a single

directly asserted relationship. Taken together, these competency questions provide a functional specification for a significant part of the TrustGuard Ontology information model. They establish what FAITH stakeholders expect to be able to retrieve without requiring those stakeholders to specify the underlying ontology structure or query language.

2.3.4 The TrustGuard Ontology Query Tool

The TrustGuard Ontology Query Tool was developed to provide a browser-based interface for querying information represented in the TrustGuard Ontology.[5] The tool enables users to formulate queries using consistent concepts without requiring knowledge of SPARQL or the underlying ontology structure. Users specify a set of query filters together with the fields that should be returned, and the application automatically generates and executes the corresponding SPARQL query against a GraphDB repository containing the TrustGuard Ontology. Query results are presented as a table within the application. The tool provides a practical mechanism for addressing the competency questions described in the preceding section. This can be illustrated using the question:

- Given a threat, which vulnerabilities are associated with it?

To answer this question, the user enters the relevant threat in the Threat field and selects the required output fields. For example, Figure 8 shows a query in which a moderately unusual threat is specified as "Oracle" and the output fields as 'Threat' and 'Vulnerability' to specify the information to be returned.

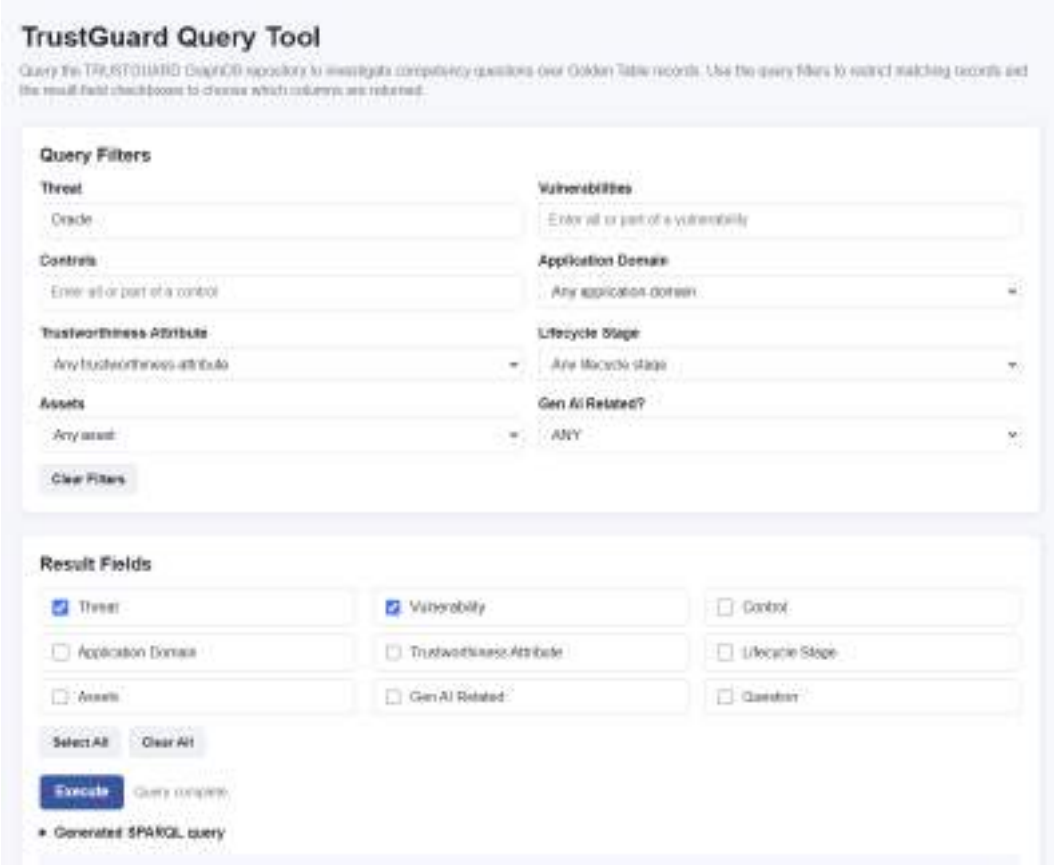


Figure 8: the TrustGuard Query Tool configured to retrieve vulnerabilities associated with an Oracle Threat

[5] The FAITH Knoweldgebase continues to evolve and is currently being extended with additional fields. The TrustGuard Query Tool described here reflects the structure of the Knowledgbase at the time of this deliverable and will require corresponding updates as new fields are incorporated into the source resource.

When the user clicks the Execute button, the application translates these selections into a SPARQL query. For this example, the generated query is:

```
PREFIX tg:  <https://git.soton.ac.uk/ontology/aito/-
/raw/main/TrustGuard.owl#>
PREFIX bfo: <http://purl.obolibrary.org/obo/>
PREFIX cco: <https://www.commoncoreontologies.org/>
PREFIX mro: <https://www.commoncoreontologies.org/mro/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?threatText ?vulnerabilityText
WHERE {
   ?record rdf:type tg:GoldenTableRecord .

   ?threatField
       rdf:type tg:ThreatFieldContent ;
       cco:ont00001761 ?threatText ;
       bfo:BFO_0000176 ?record .

   ?vulnerabilityField
       rdf:type tg:VulnerabilityFieldContent ;
       cco:ont00001761 ?vulnerabilityText ;
       bfo:BFO_0000176 ?record .

   FILTER(CONTAINS(LCASE(STR(?threatText)), LCASE("Oracle")))
}
ORDER BY LCASE(STR(?threatText)) LCASE(STR(?vulnerabilityText))
```

The query identifies Knowledge Base records containing the specified threat and the associated vulnerabilities and returns the corresponding vulnerability field contents. For this example, the results are:

Threat	**Vulnerability**
Oracle	Lack of consideration of attacks to which ML applications could be exposed to
Oracle	Lack of security process to maintain a good security level of the components of the ML application
Oracle	Poor access rights management
Oracle	The model allows private information to be retrieved
Oracle	Too much information about the model given in its outputs
Oracle	Too much information available on the model architecture
Oracle	Weak access protection mechanisms for ML model components

Table 4: Vulnerabilities associated with the Oracle Threat

The example illustrates the role played by the query tool in mediating between a stakeholder information requirement and the underlying semantic representation. The user interacts with familiar domain concepts, while the application identifies the corresponding ontology structures and constructs the required SPARQL patterns.

A second example illustrates the use of multiple result fields. Consider the competency question:

- Given a trustworthiness property, which threats and vulnerabilities are associated with it?

In this case, the user selects the required trustworthiness attribute, in this example, Explainability, as a query filter and selects both Threat and Vulnerability in the Results Fields section.

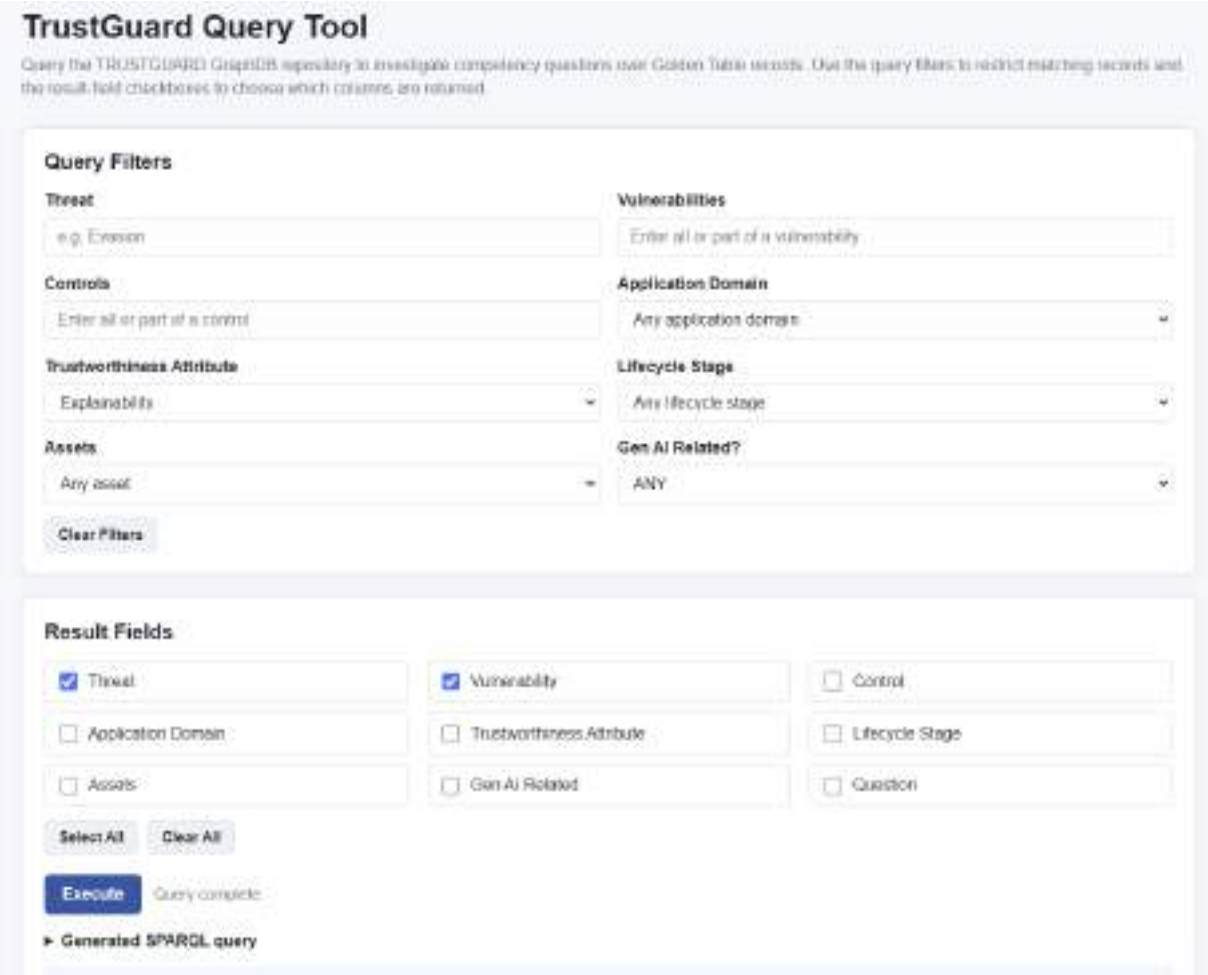


Figure 9: the TrustGuard Query Tool configured to identify which Threats and Vulnerabilities are associated with the Explainability Trustworthiness Attribute

Executing the query generates the appropriate SPARQL:

```
PREFIX tg:  <https://git.soton.ac.uk/ontology/aito/-
/raw/main/TrustGuard.owl#>
PREFIX bfo: <http://purl.obolibrary.org/obo/>
PREFIX cco: <https://www.commoncoreontologies.org/>
PREFIX mro: <https://www.commoncoreontologies.org/mro/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?threatText ?vulnerabilityText
WHERE {
   ?record rdf:type tg:GoldenTableRecord .

   ?threatField
       rdf:type tg:ThreatFieldContent ;
       cco:ont00001761 ?threatText ;
```

```
        bfo:BFO_0000176 ?record .

    ?vulnerabilityField
        rdf:type tg:VulnerabilityFieldContent ;
        cco:ont00001761 ?vulnerabilityText ;
        bfo:BFO_0000176 ?record .

     ?trustworthinessFieldContent
        bfo:BFO_0000176 ?record ;
        mro:ont00001808 ?trustworthinessValue .

    ?trustworthinessValue rdf:type bfo:BFO_0000016 .
    BIND(STRAFTER(STR(?trustworthinessValue), "#") AS ?trustworthinessLabel)

    FILTER(LCASE(STR(?trustworthinessLabel)) = LCASE("Explainability"))
}
ORDER BY LCASE(STR(?threatText)) LCASE(STR(?vulnerabilityText))
```

This returns the combinations of threats and vulnerabilities associated with the selected trustworthiness attribute.

Threat	**Vulnerability**
Failure or malfunction of ML application	Lack of explainability and traceability of decisions taken
Failure or malfunction of ML application	Variability in sensor installations and non-transparent EDA processes
Human error	Lack of documentation on the ML application
Inadequate justification for fairness metrics	Lack of detailed analysis of fairness implications
Insufficient safeguards against bias	Weakness in implementing appropriate fairness measures
Lack of critical information in training and testing data	The training and testing data are fully anonymised and lack relevant information thus hindering the correct functioning of the explanation tools
Lack of effective explanation methods	The explanation facilities have not been assessed for usability with end-users
Opaque Coverage Decisions	Lack of rationale generation in optimization logic
Over-reliance in model predictions	Over-reliance in model predictions
Poorly defined fairness thresholds	Ambiguity in fairness criteria selection
Propagation of misinformation	Absence of misinformation-specific data vetting (e.g. training or fine-tuning data is not screened for systematically false, biased, or adversarially crafted narratives, allowing corrupted content to become encoded in model weights and reproduced at inference time)
Propagation of misinformation	Lack of factual grounding and output verification mechanisms (Hallucination)
Reduced ability of operators to identify prediction errors	Loss of operator trust
Reduced model interpretability	Reduced post-hoc interpretability due lack of coupled data-predictions
Reduced operator accountability due to lack of interpretability	Operator responsibility requires explainability

Table 5: Threats and Vulnerabilities associated with the Explainability trustworthiness attribute

These examples demonstrate how the TrustGuard Query Tool operationalises the competency questions supplied by FAITH stakeholders. The distinction between query filters and result fields provides a flexible mechanism for specifying both the conditions that records must satisfy and the information that should be returned. The tool thereby provides an application layer over the TrustGuard Ontology in which user-oriented information requirements are translated into formal semantic queries and evaluated against the ontology. It also provides a real-world validation of the TrustGuard Ontology where it is evident that the ontology can competently answer plain English queries generated by non-experts.

2.3.5 TrustGuard Ontology Deployment Adaptations

The complete TrustGuard Ontology incorporates AITO together with the BFO and CCO resources on which AITO depends. The merged ontology therefore contains substantial imported content that is not directly relevant to the TrustGuard assessment tool. Hence, two additional artifacts have been produced to support a more streamlined deployment: a reduced version containing the relevant ontology modules and a version in which inferred triples have been materialised.

2.3.5.1 TrustGuard_Merged_Lite.ttl

TrustGuard_Merged_Lite.ttl provides a reduced version of the merged TrustGuard Ontology. It was generated using a custom Java utility developed using the ontology module-extraction facilities provided by the OWLAPI. All AITO and TrustGuard entities were included in the initial extraction signature, and STAR locality-based module extraction was then used to identify the subset of axioms required to preserve the relevant semantic dependencies.

This process retains the AITO and TrustGuard Ontology entities together with BFO and CCO entities and axioms included through module extraction, while removing unrelated material from the imported ontologies. Table 6 illustrates the resulting reduction.

Metric	TrustGuard Merged	TrustGuard Merged Lite
Axioms	29,757	19,124
Logical axioms	9,175	6,518
Classes	2,757	1,431
Object properties	465	80
Data properties	24	1
Individuals	1,337	1,337
Annotation properties	57	34

Table 6: Comparison of complete merged and Lite TrustGuard ontology artifacts

The Lite artifact reduces the number of axioms from 29,757 to 19,124 and the number of classes from 2,757 to 1,431. All 1,337 individuals are retained, preserving the application-specific resources represented within the TrustGuard Ontology. The particularly large reductions in classes and properties principally reflect the removal of imported BFO/CCO content that is not included in the extracted module.

The current extraction deliberately uses both AITO and TrustGuard Ontology entities as its initial signature, thereby retaining the complete AITO domain layer. A more aggressive reduction could be achieved by using only TrustGuard Ontology entities as the initial signature, allowing the extraction process to retain only those additional AITO, CCO, and BFO resources required by the resulting module.

2.3.5.2 TrustGuard_Merged_Lite_Inferred.ttl

TrustGuard_Merged_Lite_Inferred.ttl provides an alternative deployment artifact in which inferred information has been materialised. To produce this artifact, the Lite ontology was loaded into a

GraphDB repository configured with the OWL2 RL ruleset. GraphDB performed reasoning over the ontology, after which the resulting graph was exported with the inferred triples included.

The Lite and Lite_Inferred artifacts support different deployment requirements. TrustGuard_Merged_Lite.ttl retains the asserted ontology and its formal axioms, allowing a reasoning-capable system to derive relevant entailments at runtime. TrustGuard_Merged_Lite_Inferred.ttl additionally makes precomputed entailments directly available as RDF triples, which can be useful for applications that need access to inferred information without independently performing the corresponding reasoning.

Together, these artifacts provide alternative representations of the TrustGuard Ontology for downstream use: a reduced asserted ontology suitable for reasoning-enabled environments and a materialised version suitable for applications in which precomputed entailments are preferred.

2.3.6 Summary

The TrustGuard Ontology demonstrates how the reusable domain semantics provided by AITO can be specialised to support the concrete information requirements of the FAITH tool suite. The programmatic transformation of the Knowledge Base provides an ontology-based representation of project knowledge, while the stakeholder competency questions and TrustGuard Query Tool validate how this representation can support application-oriented information retrieval. The Lite and Lite_Inferred artifacts additionally provide alternative representations intended to facilitate downstream deployment.

The development of the TrustGuard Ontology also provides an application-oriented validation of AITO. Using AITO as the semantic foundation for a concrete FAITH resource tests whether its concepts and relations are sufficient to support application-level modelling and retrieval requirements. Where such requirements cannot be represented satisfactorily, they can identify potential gaps in either the TrustGuard Ontology or the underlying AITO domain model. The TrustGuard Ontology therefore serves both as an application of AITO within FAITH and as a source of requirements for its continuing refinement.

2.4 FAITH AI TAF-v2 Validation

2.4.1 Development of AI TAF-v2

AI TAF-v2 is the second iteration of the FAITH AI Trustworthiness Assessment Framework, developed under WP2 (FAITH Trustworthiness Assessment Framework Version 2) and reported in full in Deliverable D2.2 (FAITH Methodological Framework and Requirements Analysis v2). It builds directly on TAF-v1, reported in D2.1, rather than replacing it: the six-phase structure (Cartography, Threat Analysis, Impact Assessment, Vulnerability Analysis, Risk Analysis, Countermeasures), the human-centric approach, and the alignment with ENISA, NIST, and ISO risk-management and AI-governance standards are all retained unchanged from TAF-v1.

The development of TAF-v2 was motivated by limitations identified in TAF-v1's in-isolation approach to assessing individual AI assets (data, models, humans) separately from the broader socio-technical ecosystem in which they operate. As documented in D2.2, this approach does not account for an asset's network of relationships with other software, infrastructure, people, places, and jurisdictions; does not reflect the specific real-world context of deployment; and does not capture cascading effects as threats, consequences, and impacts propagated across connected assets and systems. As a result, in-isolation assessment alone can lead to incomplete risk identification, unrealistic trustworthiness

conclusions, and mitigation recommendations that patch individual assets rather than address system-wide risk.

To address these limitations, TAF-v2 introduces seven key extensions to TAF-v1, all developed and documented in D2.2:

- A shift from assessing isolated assets to assessing ecosystems of connected assets, spanning software, infrastructure, people, places, and jurisdictions (D2.2, Chapter 3).
- Explicit modelling of cascading and propagating risks across systems (D2.2, Section 3.4).
- Integration of in-isolation assessment outputs as fine-tuning inputs into context-specific, in-context system modelling (D2.2, Sections 4.1.3–4.1.4).
- A more detailed classification of AI model types and their associated threats and controls, with particular attention to LLM and Agentic AI systems relevant to the FAITH LSPs (D2.2, Section 3.5).
- The introduction of system-level Control Strategies, rather than individual asset-level controls (D2.2, Section 4.1.4.3).
- A move from single-dimension trustworthiness optimisation to balanced, evidence-based trade-off management across trustworthiness attributes and operational constraints, informing risk tolerance decisions (D2.2, Section 3.7).
- Refined and redefined steps for each of the six TAF phases, with an added requirement to document the actions taken at each step (D2.2, Chapter 4).

TAF-v2 was developed through a structured, multi-disciplinary process coordinated across the WP2 consortium. As part of this same development process, a validation plan for TAF-v2 was also established, targeting implementation throughout the project duration. This plan directly informed the expert validation workshops reported in Sections 2.2.1–2.2.5 of this deliverable, which assess the credibility, usefulness, and usability of the resulting TAF-v2 framework.

2.4.2 Methodology

TAF-v2 was validated through two structured expert workshops, rather than a single formal survey instrument, reflecting the more conceptual and cross-role nature of the framework compared to the individual FAITH tools. Both workshops followed the same structure: a short overview presentation of the TAF-v2 six-phase model (Cartography, Threat Analysis, Impact Assessment, Vulnerability Analysis, Risk Analysis, and Countermeasures) and its extension from the in-isolation, single-asset assessment of TAF v1 to the in-context, whole-system assessment introduced in TAF-v2; an individual online survey capturing initial thoughts; and a facilitated group discussion, recorded and transcribed.

As set out in D2.2, this validation activity assesses TAF-v2 as a method only, not the FAITH tools that implement it; tool-level validation is reported separately. D2.2's original validation plan specified three separate 60-minute online workshops, one per stakeholder group (developers team, risk assessor group, end user group), using adapted Technology Acceptance Model (TAM) criteria. In delivery, this was consolidated into two mixed-role workshops covering all three stakeholder groups together (22 July and 3 August 2026), with role-based breakout discussion used in the first session to preserve some separation of perspective.

The first workshop was held on 22 July 2026 with a broad consortium audience drawn from developer, risk assessor, and end user roles across multiple LSPs (Media, Transport, Education, Underwater, Wastewater, Healthcare, and Active Ageing). Participants first completed an individual "Initial Thoughts" survey addressing three areas: conceptual clarity, completeness, and internal coherence; understandability, usefulness, and relevance to their role; and integration potential, adoption likelihood, and support for accountability. Discussion then moved into breakout sessions organised by

role (developer, risk assessor, end user), with group discussion feedback captured through a second, structured "Group Feedback" survey.

The second workshop was held on 3 August 2026 with the Norwegian LSP team, consisting of a developer, a risk assessor, and an end user. Given the smaller group size, this session ran without breakout rooms, with discussion held as a single group and feedback recorded using the same "Initial Thoughts" and "Group Feedback" survey instruments used in the first workshop.

Across both workshops, discussion was organised around three validation dimensions: credibility (clarity, completeness, internal coherence, and domain coverage), usefulness (understandability, explainability, and relevance to the participant's role), and usability (potential for integration into existing workflows, likelihood of adoption, and support for accountability and documentation). Each workshop closed with a structured consensus exercise: participants rated TAF-v2 against a four-point scale (*validated as-is; validated with minor refinements; partially validated, substantial refinements required; not validated for current purpose*), nominated the framework's top three strengths and weaknesses, and answered the overarching validation question — "is the Trustworthiness Assessment Framework Version 2 a credible, usable, and sufficiently useful framework for AI trustworthiness assessment in real-world contexts?" — with a consensus "broadly yes" or "basically no."

2.4.3 TAF-v2 Validation Results

The Initial Thoughts survey drew responses from participants across both workshops, spanning developer, risk assessor, and end-user perspectives from LSPs including Media, Transport, Education, Underwater Robotics, Wastewater, Healthcare, Ports, and Active Ageing. Conceptual clarity was rated positively overall: most respondents described the framework as clear, complete, and internally coherent, and several developers and risk assessors highlighted the logical progression from Cartography through Threat Analysis, Impact Assessment, Vulnerability Analysis, Risk Analysis and Countermeasures, the coverage of technical, human, organisational, legal, and social dimensions, and the seven NIST-aligned Trustworthiness Attributes (TWAs) as key strengths. A recurring caveat concerned terminology: several respondents, particularly end users and those without a security or risk-management background, noted that terms such as asset, control, and vulnerability required more explanation, and that the boundary between the Threat, Vulnerability, Impact, and Risk phases was not always intuitive.

On understandability and usefulness, responses diverged by role. Developers and risk assessors generally found the six-phase structure understandable and highly relevant to their work. End users, and organisations not yet using AI operationally, found the framework more abstract and reported needing domain-specific worked examples before they could confidently explain it to others. Several respondents noted that TAF-v2 would increase their confidence in an AI system that had been assessed under the framework, even without being able to apply it themselves.

Perceptions of integration and adoption were the most conditional of the three dimensions. Respondents generally saw good potential for integration into existing quality, governance, and compliance processes, but consistently flagged that:

1. the full framework capability — cascading threats and whole-system, in-context risk assessment — could not yet be exercised, because the SSM tool underpinning it was still under development; and
2. integration required an explicit mapping of what TAF-v2 covers relative to existing organisational quality and compliance processes, to avoid duplication. Several respondents not currently using AI operationally rated adoption likelihood as conditional on future AI adoption rather than a present-day decision.

Group discussion, captured separately from the individual survey, surfaced more specific and often more critical feedback. Common threads across both workshops included:

- **Phase-to-phase rationale.** Participants — notably the risk assessor from the Norwegian LSP — noted that although the six phases "seem to cover the ground" needed for a trustworthiness risk assessment, it was "not evident throughout how the different phases build on each other," and this could be made more explicit.
- **Framework–tool conflation.** Several participants, especially those already using TAF v1 through TrustGuard, found it difficult to separate the abstract framework from its tool implementation, and noted that TAF-v2's in-context capabilities were hard to judge before the SSM tool became available.
- **Duplication with existing processes.** A request for a structured account of what TAF-v2 specifically covers that is not already addressed by existing organisational policies, standards, and quality processes, together with how each phase maps to those existing structures — raised most explicitly regarding the legislative and standards mapping step of Cartography, which one participant noted already requires substantial specialist effort under TAF v1.
- **Worked examples.** Requests for concrete, domain-specific worked examples (health, press/media, underwater robotics, ports) and at least one fully worked quantitative example showing how a cascading Impact Level would be calculated end-to-end across a chain of connected assets, since the current description of cascading impact is narrative rather than demonstrated.
- **Formalising the TrustGuard–SSM dependency.** A suggestion to document the TrustGuard–SSM integration as a named project dependency, since it currently underpins TAF-v2's central in-context claim without being a contracted task in its own right.
- **Validation consistency.** A proposal to strengthen the validation methodology with an inter-rater consistency check — two assessors independently applying TAF-v2 to the same scenario — since the current approach tests usability rather than whether the framework produces consistent results across users.
- **Auto-populating asset metadata.** Suggestions to auto-populate asset metadata (model type, learning algorithm, function, data type) into the Cartography phase directly from the AI Model Passport and Data Passport, to reduce what several respondents identified as the most labour-intensive step of the framework.
- **Documentation templates for system-level phases.** A request for a dedicated documentation template or checklist for the system-level, cascading-threat phases, mirroring the structured tables already used for asset-level TWA metrics.
- **Uneven documentation of complementary metrics.** Clarification requests on whether Green AI/Sustainability and Uncertainty/Calibration should be treated as core TWAs on a par with the other seven, or as a supplementary layer, since D2.2's rationale for their treatment was not evidently surfaced clearly enough in the workshop overview materials.

Despite these refinement requests, sentiment toward the overall validation statement was consistently positive: end user participants confirmed that knowing an AI system had gone through the TAF-v2 process would increase their confidence in it, and developer and risk assessor participants generally agreed the framework was internally consistent, even allowing that inconsistencies might only surface once applied in practice over time.

Both workshops closed with a structured consensus exercise, using a four-point scale — validated as-is; validated with minor refinements; partially validated, substantial refinements required; or not validated for current purpose. In the 22 July workshop, live polling of the group converged clearly on "validated with minor refinements," explicitly ruling out the more severe "substantial refinements"

option. The group then answered the overarching validation question with a consensus of "broadly yes."

In the 3 August (Norwegian LSP) workshop, one participant initially suggested "validated as-is" was viable, but the group converged on the same "minor refinements" position, tying this explicitly to the phase-to-phase rationale clarity raised earlier in discussion (see Finding 3). No participant in either workshop supported "partially validated" or "not validated."

2.4.4 Key Actionable Findings

Finding 1 — Strong perceived credibility, conditional usability. Across both workshops, respondents rated TAF-v2 as conceptually clear, complete, and internally coherent, and consistently confirmed that a TAF-v2-assessed system would carry greater credibility with end users. However, usability ratings were consistently qualified by the framework's dependence on tooling (SSM) not yet available, meaning respondents could evaluate the in-isolation, asset-level half of TAF-v2 far more confidently than the in-context, whole-system half.

Finding 2 — Framework–tool mix-up obscures independent evaluation of TAF-v2. Several risk assessors and developers, particularly those already using TrustGuard under TAF v1, found it difficult to separate their assessment of the TAF-v2 method from their experience of its planned tool implementation. This blurs validation of the framework as a method independent of its tooling and should be addressed explicitly in future validation rounds once SSM is available.

Finding 3 — Phase-to-phase rationale needs to be made explicit. Multiple respondents pointed out that while the six phases cover the necessary ground, the logic connecting one phase to the next is not made sufficiently explicit in current TAF-v2 documentation.

Finding 4 — Terminology and abstraction level create an accessibility barrier for non-specialists. End users and respondents from organisations with limited AI experience consistently found core terms (asset, control, vulnerability, TWA) unfamiliar, and reported needing simplified language, visual aids, and worked examples before being able to explain the framework to others.

Finding 5 — Worked, domain-specific examples are the most requested improvement. Across nearly every respondent group (health, press/media, underwater robotics, ports, transportation), the single most consistent request was for concrete examples showing how each phase — particularly the cascading-impact calculation — is applied in practice, rather than described narratively.

Finding 6 — Duplication with existing organisational processes is a barrier to adoption. Several respondents noted that organisations already operate quality, compliance, and standards processes, and requested an explicit account of what TAF-v2 covers that is genuinely new, to avoid redundant effort — most acutely at the legislative/standards mapping step of Cartography. Separately, the suggestion to formalise the TrustGuard–SSM integration as a named dependency, raised in group discussion, is substantiated by D2.2 itself, which states that this integration — automatically ingesting asset-level Impact Levels per Trustworthiness Attribute and Controls from TrustGuard into the System Security Modeller (SSM) — is not explicitly detailed in the Grant Agreement as a Task, despite being treated by the WP2 team as vital to TAF-v2.

Finding 7 — Cartography is the most labour-intensive entry point. Both workshops identified asset characterisation at the Cartography stage as the most burdensome part of the framework to complete manually, and proposed auto-populating this from the AI Model Passport and Data Passport.

Finding 8 — Documentation guidance is uneven across phases. The asset-level, in-isolation phases benefit from structured, checklist-style guidance, whereas the system-level, cascading-threat phases

are described narratively, making it harder to know when a whole-system trustworthiness assessment is complete.

Finding 9 — The validation methodology does not yet test consistency across assessors. Current validation activity primarily tests usability and acceptance (aligned to TAM), rather than whether independent risk assessors would reach consistent conclusions when applying TAF-v2 to the same scenario.

Finding 10 — Uneven documentation of complementary metrics outside the seven core TWAs. D2.2 already gives an explicit rationale for Green AI/Sustainability metrics: they are deliberately not mapped to a single Trustworthiness Attribute, but treated as complementary to Safe AI in a systemic, lifecycle sense (D2.2, Section 2.6.6). This rationale was evidently not surfaced clearly enough in the workshop overview materials, since participants raised it as an open question. Uncertainty/Calibration metrics (D2.2, Section 2.6.7), by contrast, have no equivalent explicit rationale in current documentation — this is a genuine gap rather than a communication issue.

2.4.5 Updates

Based on the findings above, the following updates to TAF-v2 and its supporting documentation have been identified:

Planned for TAF-v2 documentation:

- An explicit narrative connecting each phase to the next, addressing Finding 3.
- A glossary of core terms and simplified, non-specialist explanatory text for each phase, addressing Finding 4.
- One or more fully worked, domain-specific examples — including a complete quantitative worked example of a cascading Impact Level calculation — addressing Finding 5. Candidate domains include healthcare, underwater robotics, and press/media, reflecting the LSPs that raised this need most strongly.
- A dedicated documentation template/checklist for the system-level, cascading-threat phases, addressing Finding 8.
- Surface the existing D2.2 rationale for Green AI/Sustainability's relationship to Safe AI more prominently in TAF-v2 overview/workshop materials, and add an equivalent explicit rationale for Uncertainty/Calibration's status relative to the seven core TWAs, addressing Finding 10.

Planned for tooling and process (dependent on SSM release):

- Auto-population of Cartography-stage asset metadata from the AI Model Passport and Data Passport, addressing Finding 7.
- Document the TrustGuard–SSM integration as a formally tracked task/dependency rather than an informal WP2 team decision, addressing Finding 6, supported directly by D2.2's own acknowledgement that this integration is not currently a Grant Agreement Task.

Planned for future validation activity:

- An inter-rater consistency check, in which two or more assessors independently apply TAF-v2 to the same scenario, addressing Finding 9.
- A follow-up validation round once the SSM tool is released, to allow the in-context, whole-system half of TAF-v2 to be assessed directly, addressing Findings 1 and 2.

TAF-v2 is being operationalised across the FAITH tool ecosystem in two stages, reflecting its extension of the in-isolation assessment introduced in TAF v1 into the in-context, whole-system assessment new to TAF-v2.

The in-isolation half of TAF-v2 — asset classification (model architecture, learning algorithm, and function), human actor identification, and asset-level threat, vulnerability, and risk analysis — builds directly on the existing TrustGuard (for AI models and datasets) and TrustSense (for human/team maturity) tools already used under TAF v1.

The in-context, whole-system half of TAF-v2 is implemented through integration between TrustGuard and the System Security Modeller (SSM). Per D2.2, this integration allows SSM to automatically ingest asset-level Impact Levels captured in TrustGuard — organised by Trustworthiness Attribute — to fine-tune the corresponding Impact Levels in the whole-system model, and to ingest information on asset-level Controls so that SSM can automatically mark those Controls as present at system level. D2.2 explicitly notes that although this integration is not detailed in the Grant Agreement as a formal Task, the WP2 team consider it vital to a nuanced, contextual implementation of TAF-v2, and that further technical integration and Knowledge Base-mapping work is ongoing, to be reported in D2.4.

At the time of the validation workshops (July–August 2026), the updated TrustGuard and SSM tools had not yet been made available to LSPs; the workshop facilitator indicated a target release from around November 2026. This means validation participants were only able to evaluate TAF-v2's in-context capabilities as a specification rather than a deployed tool, a dependency reflected in Findings 1 and 2 above. Once released, TrustGuard v2 and SSM together are intended to give risk assessors an end-to-end path through TAF-v2: TrustGuard and TrustSense producing in-isolation asset- and human-level assessments, which then feed as structured inputs into an SSM-based in-context system model, from which cascading risk, suggested control strategies, and trade-off decisions across Trustworthiness Attributes can be derived.

2.4.6 Summary of the validation outcomes

Across both validation workshops, TAF-v2 achieved a consistent expert consensus of validated with minor refinements. Using the four-point validation scale presented to participants (validated as-is; validated with minor refinements; partially validated with substantial refinements required; not validated for current purpose), both the 22 July and 3 August sessions converged independently on the same rating, and both groups answered "broadly yes" to the overarching question of whether TAF-v2 is a credible, usable, and sufficiently useful framework for AI trustworthiness assessment in real-world contexts. No respondent group selected "partially validated" or "not validated" at either workshop.

This outcome is corroborated by the individual Initial Thoughts survey and group discussion data reported in Sections 2.2.2 and 2.2.3: conceptual clarity, completeness, and internal coherence were rated positively across roles and LSPs, and the framework's core structure (the six phases, the seven NIST-aligned TWAs, and the in-isolation-to-in-context extension) was not challenged by any respondent. The refinements requested were consistently the kind that would sharpen and operationalise the framework rather than redesign it: making the phase-to-phase logic explicit (Finding 3), reducing terminology and abstraction barriers for non-specialists (Finding 4), and adding worked, domain-specific examples (Finding 5) — none of which questions TAF-v2's underlying validity.

The one caveat to this positive rating is scope: as recorded in Findings 1 and 2, participants could only meaningfully evaluate the in-isolation, asset-level half of TAF-v2 first-hand, since the SSM tool needed to exercise the in-context, whole-system half was not yet available at the time of either workshop. The "minor refinements" rating should therefore be read as applying most confidently to the framework's asset-level half, with the in-context half provisionally accepted on the strength of its documented design (D2.2) rather than through direct hands-on evaluation.

Taken together, the evidence supports an overall validation rating of credible and usable, with minor refinements required before wider rollout — specifically the documentation, terminology, and worked-example improvements identified in Section 2.2.4, plus a follow-up validation round once the SSM/TrustGuard v2 tools are released to directly test the in-context capabilities that could not be assessed in this round.

3. FAITH IN-ISOLATION TRUSTWORTHINESS ASSESSMENT TOOLS VALIDATION

3.1 Methodology

The three core in-isolation trustworthiness assessment tools - AI Model Hub, TrustSense, and TrustGuard - were deployed during the Phase 1 Pilots within each of the 7 LSPs. A standardised and formalised feedback process was developed, consisting of formal, standardised, online mixed-method feedback survey adapted for each tool.

The surveys were designed in accordance with the established Technology Acceptance Model (versions 1, 2 and 3) (TAM) measures of Usability, Usefulness, and User Perceptions & Attitudes. In addition, and in line with D3.3, a section within each survey containing the relevant NIST Measures was also included for each tool. Hence, the survey consisted of 6 sections and 27-30 questions:

1. About You (5 closed questions)
2. Perceived Usefulness (4 multiple choice questions + 1 open question)
3. Ease of Use (4 multiple choice questions + 2 open questions)
4. Human Factors and Social / Organisational Influence (7 multiple choice questions)
5. NIST MEASURES (multiple choice questions)
6. Detailed Suggestions (2 open questions)

The surveys were tested with colleagues, which led to several improvements in question wording. The surveys were then published on the University of Southampton's Qualtrics online survey platform.

The surveys were standardised (with minor tweaks for relevance to each tool) in order that the results could be reasonably compared across the three tools and the findings applied at both the individual tool level and across the FAITH tool suite as a whole.

During the training workshops for the LSP's risk assessors on how to use each of the tools, the requirement to complete the formal feedback after using the tools was emphasised. Additional informal feedback opportunities were also provided during WP2 weekly meetings and by emails from LSP partners and developer partners directly to the tool providers.

In addition to the standardised Qualtrics surveys deployed across the Large Scale Pilots (LSPs) for the three in-isolation tools, trustilio established dedicated, multi-modal feedback loops to validate the trustSense maturity assessment instrument. Given that trustSense serves to evaluate the readiness and capabilities of human teams across the AI lifecycle, the validation approach required channels that went beyond isolated, individual responses.

To achieve this, trustilio utilised the following methodology:

- **Facilitated Consensus-Building Workshops:** Prior to final deployment, trustSense was evaluated through structured physical and hybrid workshop settings, notably the 1st FAITH Workshop. These sessions gathered project partners, domain experts, cybersecurity professionals, and end-users to explicitly cross-examine the clarity of the assessment questions, the consistency of the proposed scales, and the socio-technical dynamics of team scoring. This workshop environment initiated a methodological shift from individual response collection to team-based, consensus-driven profiling, ensuring the tool aligns with the collective decision-making structures of complex organisations.
- **Simulated Evaluation and Sensitivity Runs:** The tool was subjected to hundreds of simulated evaluation runs and applied across diverse hypothetical organisational scenarios. This automated

testing channel allowed trustilio to analyse score behaviour, trace response patterns, and evaluate the sensitivity of the algorithm to varying team profiles, ensuring operational robustness before pilot entry.

- **Targeted Controlled Group Deployments:** Structured assessment runs were conducted with university student cohorts. This auxiliary feedback channel was explicitly used to stress-test question clarity, isolate abstract constructs (such as responsibility distribution and socio-technical interactions), and measure internal consistency in a controlled environment.
- **Iterative Post-Workshop Digital Tracking:** Following interactive workshop sessions, participants were granted access to online testing surveys to critically review draft scales, provide localised notes on question wording, and balance technical depth with accessibility.

Structured inputs were supplemented by iterative informal feedback collected during dedicated workshops, facilitated discussions and exchanges via the online chat function. This process allowed participants to provide real-time comments on the clarity, relevance, usability and sectoral applicability of the tool, identify areas requiring refinement, and discuss the alignment of the assessment and resulting recommendations with the operational realities of their respective pilots. Although some of these activities did not involve FAITH tool users, collectively, the formal and informal evidence collected using this methodology, together with the formal FAITH survey feedback, provided a more comprehensive basis for refining the TrustSense tool than the survey alone.

3.2 AI Model Hub and Passport

The FAITH AI Model Hub is a web-based platform developed throughout the FAITH pipeline to enable systematic registration, storage, and management of AI model and dataset metadata through the context of Data Passports and AI Model Passports across the FAITH Large-Scale Pilots (LSPs). The platform supports risk assessors, developers, and end-users in documenting the provenance, configuration, and deployment context of AI systems under assessment, thereby contributing to the transparency and accountability requirements of the FAITH Trustworthiness Assessment Framework (AI_TAF). The feedback reported below was collected through a structured formal survey administered to LSP participants following the initial utilisation of the platform, and is supplemented by informal feedback gathered through partner interactions and technical support exchanges.

3.2.1: Formal & Informal LSP Feedback Results

A formal feedback survey was distributed to representatives from all seven LSPs. In total, thirteen (13) valid responses were received from participants across

- LSP 1 (Media),
- LSP 2 (Transportation),
- LSP 3 (Education),
- LSP 4 (Underwater Drones),
- LSP 5 (Wastewater Management),
- LSP 6 (Healthcare), and
- LSP 7 (Active Ageing).

Respondents represented organisations including FORTH, ATC, SINTEF, ICCS, UNIGI, UNIFI, and BRIDG. The survey was structured into five thematic sections, as detailed below.

Section 1: About You

The survey opened with demographic and contextual questions to characterise the respondent sample and ensure appropriate interpretation of subsequent findings.

Table 7 Summary of demographic and contextual characteristics of survey respondents (N=13, excluding one test entry). Counts marked with ~ are approximate, read from bar charts in the original survey report.

Question	Key Findings
Q1 – LSP represented	Responses were obtained across all 7 LSPs. LSP 5 (Wastewater) is most represented (x4), followed by LSP 1 (Media) (x2) and LSP 6 (Healthcare) (x3). LSP 4 (2 responses), LSP 2, 3, and 7 each contributed x1 response.
Q2 – Organisation represented	12 responses from 7 organisations: SINTEF (×3), ATC (×3), FORTH (×2), UNIFI (×2), ICCS (×1), BRIDG (×1). One test entry was excluded from analysis.
Q3 – Primary role	Risk Assessors and Developers were equally represented (x6 each), with a smaller number of End Users (x2), indicating a technically experienced respondent pool.
Q4 – Lifecycle phase	The majority of respondents' AI models are in the Deployment Phase (x10), with fewer in Development (x4) or Design (x1) phases, indicating real-world operational context.
Q5 – Deployment environment	Most models are in full deployment in a real environment (x6), with a smaller number in sandbox/testing environments (x2) or other configurations (x2).

Figure 10 summarises these distributions graphically across all five demographic variables.

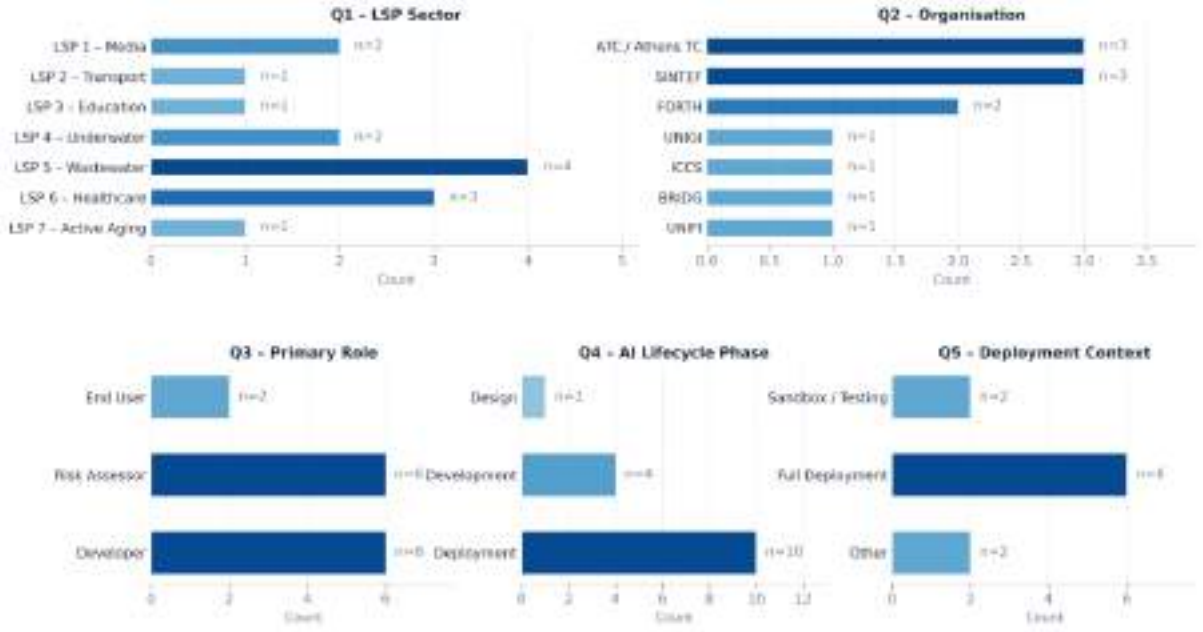


Figure 10 Demographic profile of the survey respondents. Top row: distribution by LSP sector and organisation. Bottom row: primary role, AI lifecycle phase, and deployment environment.

Section 2: Perceived Usefulness (link with Technology Acceptance Model (TAM) 1)

The Perceived Usefulness was assessed through four Likert-scale items (Q6–Q9) which examine whether the AI Model Hub improves the transparency, productivity, efficiency, and overall usefulness in risk assessment. The results indicate a generally positive reception, with a dimension mean of μ =

3.68 / 5.0. The highest-rated item was Q6 (Improves transparency of risk assessment work, μ = 3.83), with the majority of respondents selecting "Somewhat Agree" or "Strongly Agree". Q9 (Useful for risk assessment work overall, μ = 3.75) also returned a strong score. Q8 (More efficient than usual methods, μ = 3.50) was the most cautious, with a higher proportion of neutral responses, suggesting users see value in the tool but are not yet experiencing strong efficiency gains relative to their existing workflows.

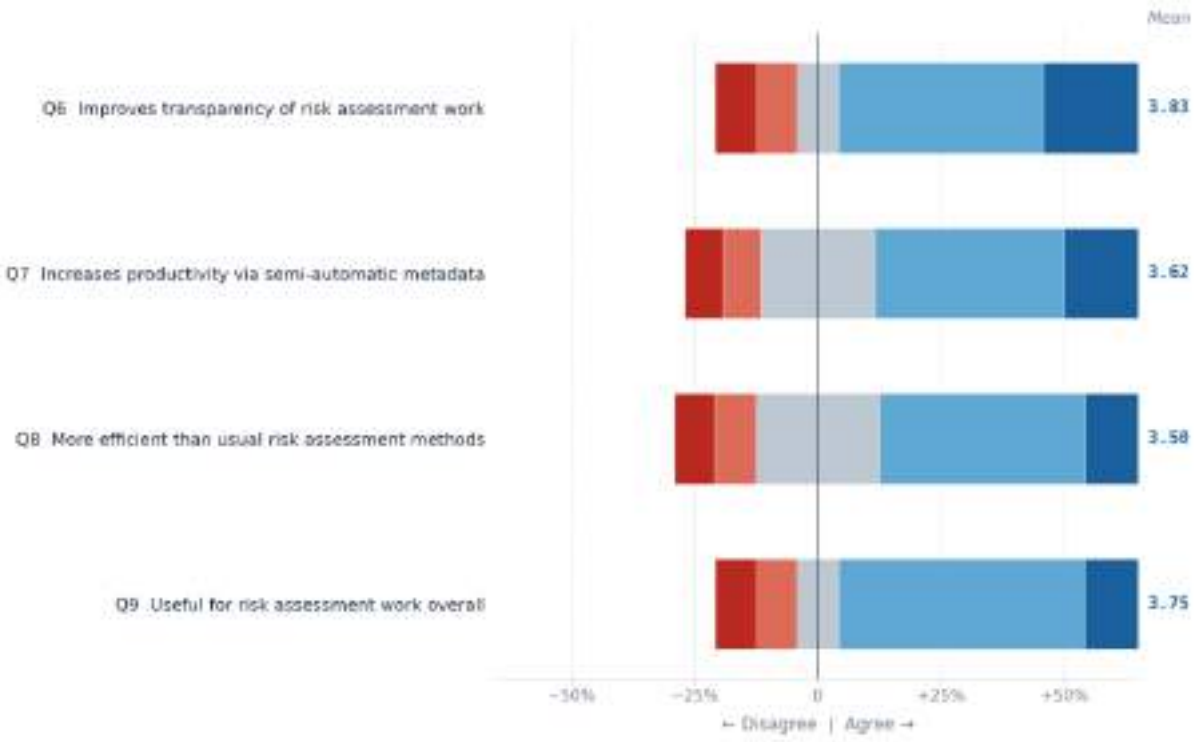


Figure 11 Response distributions and mean scores for Perceived Usefulness items (Q6–Q9). Bars represent percentage of respondents per response category.

Section 3: Ease of Use (link with TAM 1)

The Perceived Ease of Use was measured through four items (Q11–Q14) addressing learnability, intuitiveness, technical accessibility, and overall ease of use. The dimension mean was μ = 3.59 / 5.0. Q12 (Interaction is intuitive, clear, and understandable, μ = 3.83) was the strongest result in this section, indicating that the interface itself is well-received. Q11 (AI Model Hub is easy to learn, μ = 3.64) and Q14 (Easy to use overall, μ = 3.64) also returned positive scores. The most notable result was Q13 (Does not require high levels of technical knowledge, μ = 3.23), which produced the most divided response distribution in this section, with a meaningful proportion of respondents selecting "Somewhat Disagree" or "Strongly Disagree". This suggests that while the interface is intuitive for technically experienced users, non-technical participants, particularly risk assessors, encounter a steeper learning curve. Open-text responses (Q10, Q15) confirm this, requesting in-context guidance and tooltips within the AI model Hub.

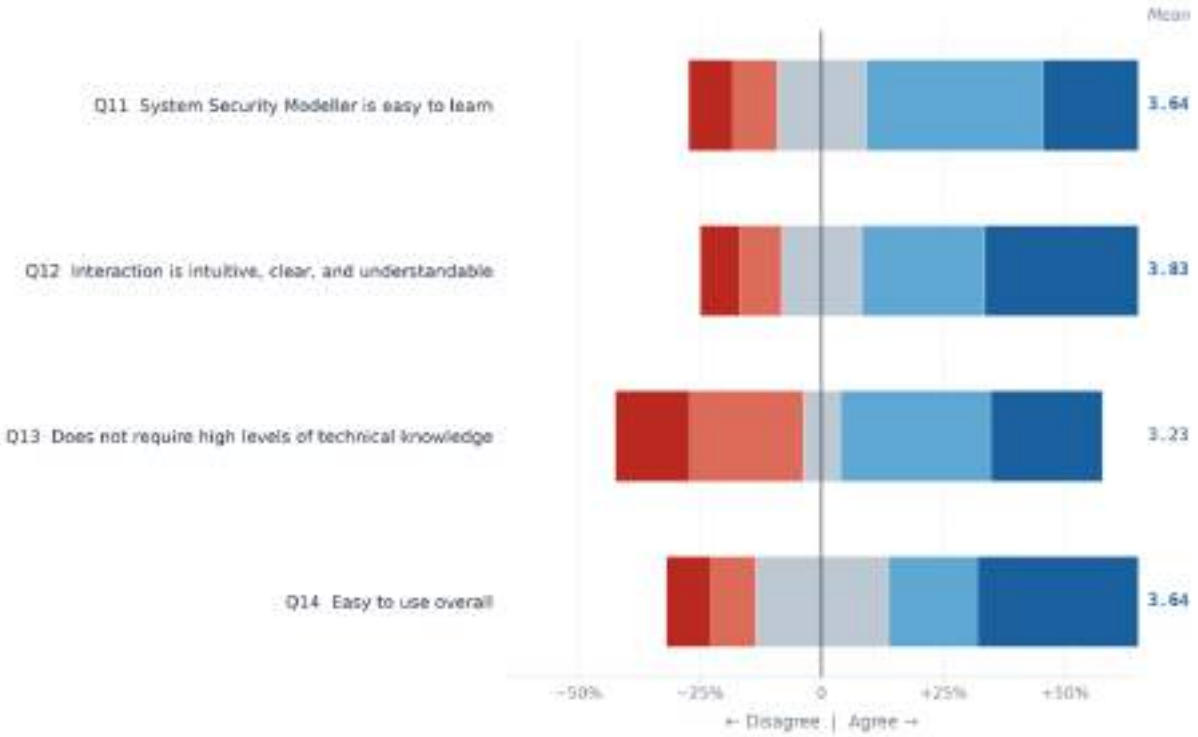


Figure 12 Response distributions and mean scores for Ease of Use items (Q11–Q14).

Section 4: Human Factors (link with TAMs 2 & 3)

Trust & Confidence (Q16–Q17). Respondents expressed a generally positive level of trust in the AI Model Hub, with a sub-dimension mean of μ = 3.68 / 5.0. Q17 (Trust the outputs and recommendations produced, μ = 3.71) slightly outperformed Q16 (Confident in ability to use the Hub correctly, μ = 3.64), though both items showed a notable proportion of neutral responses (~29%), reflecting early-stage familiarity with the tool.

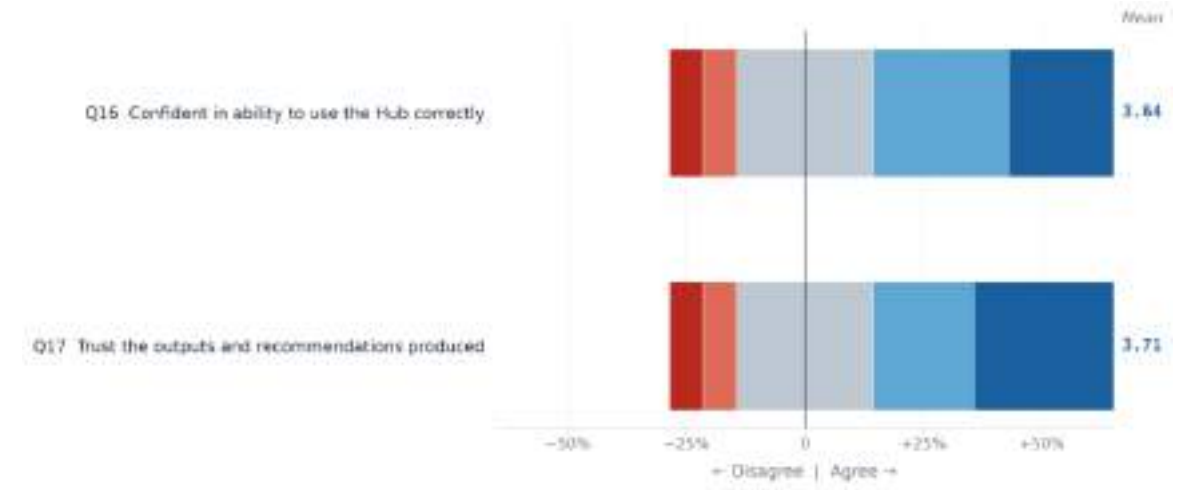


Figure 13 Respondent ratings for the Trust & Confidence dimension of the FAITH AI Model Hub. [6]

[6] Bars show the distribution of responses from Strongly Disagree (red, left) to Strongly Agree (blue, right); the centre line represents the neutral boundary. Mean scores are annotated to the right of each item (blue: ≥3.50; grey: 3.0–3.49; red: <3.0).

Attitude & Social Influence (Q18–Q20). Overall attitude toward the Hub (Q18, μ = 3.45) was moderately positive. However, social influence indicators were weaker: Q19 (Colleagues or peers are impressed by the Hub, μ = 3.09) returned the lowest score of this section, with the majority of responses concentrated in the neutral category, suggesting limited peer visibility of the tool at this stage. Q20 (Senior staff or management support its use, μ = 3.27) was similarly neutral, indicating that organisational endorsement has not yet been strongly established across all LSPs.

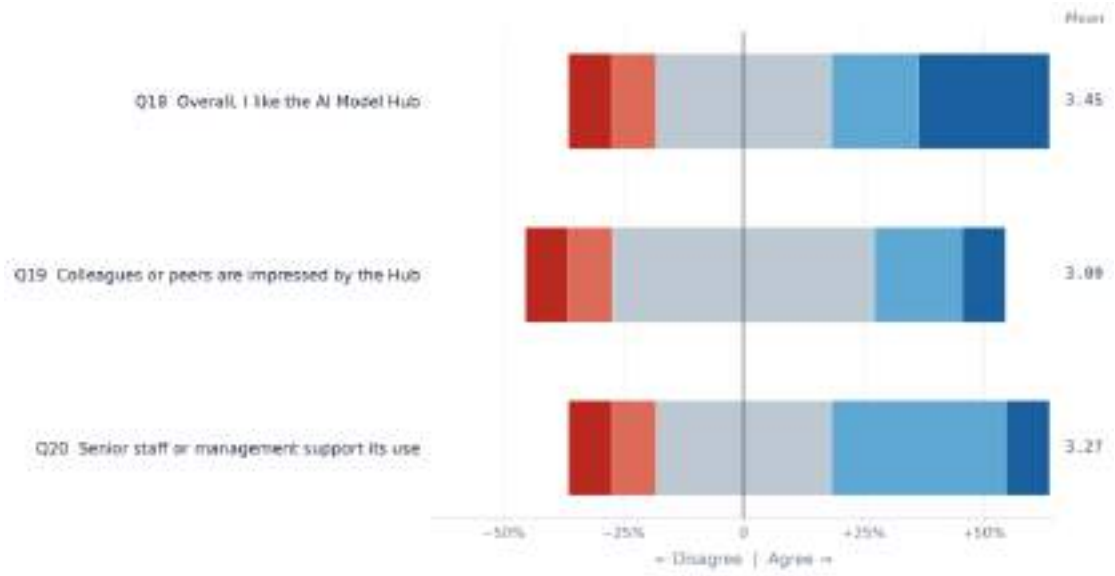


Figure 14 Respondent ratings for the Attitude & Social Influence dimension of the FAITH AI Model Hub. Items assess overall satisfaction, willingness to recommend the platform to colleagues, and perceived organisational benefit.

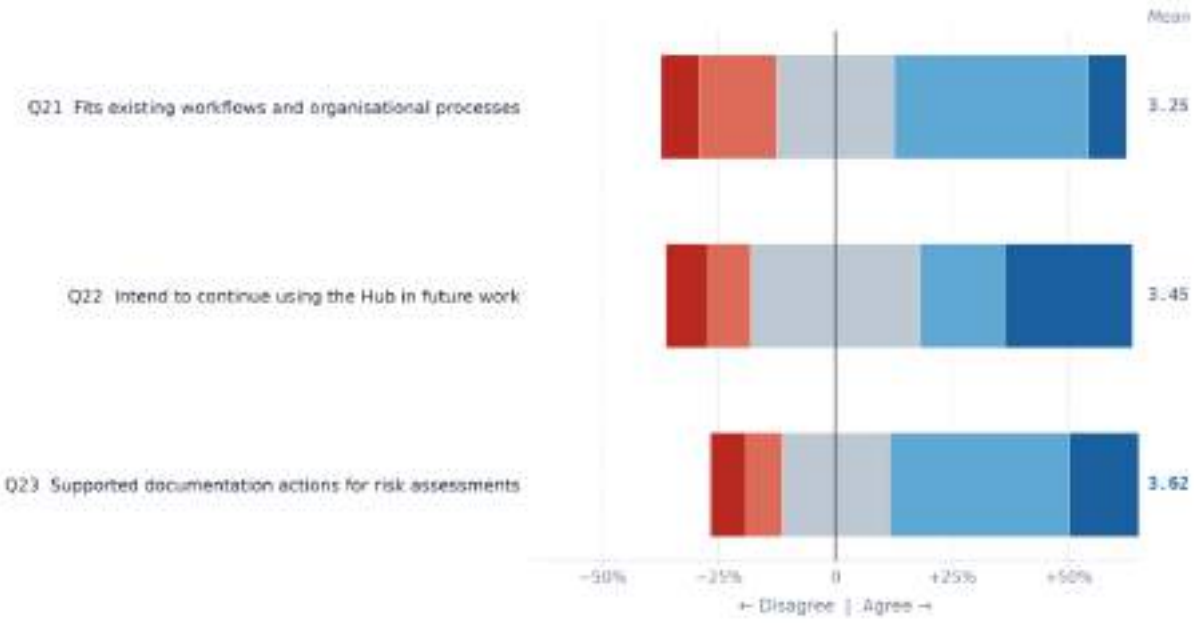


Figure 15 Respondent ratings for the Compatibility & Adoption dimension of the FAITH AI Model Hub. Items address integration with existing workflows, fit within current toolchains, and intention to continue using the platform.

Compatibility & Behavioural Intention (Q21–Q22). Q21 (Fits existing workflows and organisational processes, μ = 3.25) was the weakest item in this group, with a notable share of "Somewhat Disagree" responses, indicating that integration into day-to-day practices remains a challenge for some LSPs.

Q22 (Intend to continue using the Hub in future work, μ = 3.45) was more positive, suggesting willingness to adopt the tool despite current friction points.

Across all Human Factors dimensions, the pattern is consistent with early-stage adoption: users are cautiously positive but have not yet reached the level of fluency and organisational embedding that would produce stronger scores on social influence and compatibility measures.

Section 5: NIST Measures

Two items (Q24–Q25) addressed the AI Model Hub's support for AI trustworthiness requirements aligned with the NIST AI RMF MEASURE function, as defined and mapped in D3.3. Specifically, Q24 maps to Measure 2.8 (Risks associated with transparency — FAITH TWA: Accountability & Transparency) and Q25 maps to Measure 2.9 (The AI model is explained, validated, and documented — FAITH TWA: Explainability & Interpretability).

Q24 (Supports transparency requirements through data and model metadata, μ = 3.82) was one of the highest-scoring items across the entire survey, with the majority of respondents selecting "Somewhat Agree" or "Strongly Agree". This indicates that the Hub's metadata management and documentation functions are perceived as meaningfully supporting Measure 2.8, consistent with D3.3's identification of the AI Model Hub, AI Model Passport, and Data Passport as primary tools for transparency and accountability evidence provision.

In contrast, Q25 (Supports explainability requirements through data and model metadata, μ = 2.90) was the lowest-scoring item in the entire survey and the only item falling below the neutral midpoint. The response distribution was markedly negative, with "Somewhat Disagree" being the modal response. This points to a specific and significant gap in Measure 2.9 compliance. While D3.3 identifies the AI Model Hub as a key enabler of explainability through model cards and structured metadata, the survey results indicate that users do not yet perceive this as sufficient to support the interpretability and reasoning-transparency dimension of Measure 2.9 — that is, documenting why an AI model behaves as it does, not only what it is and how it was trained. This gap is directly relevant to NIST and EU AI Act explainability obligations and should be treated as a development priority. A related structural limitation was also noted for use cases built on pre-trained LLMs, where the current model card schema does not adequately capture the relevant metadata, further constraining explainability support for this increasingly common deployment pattern.

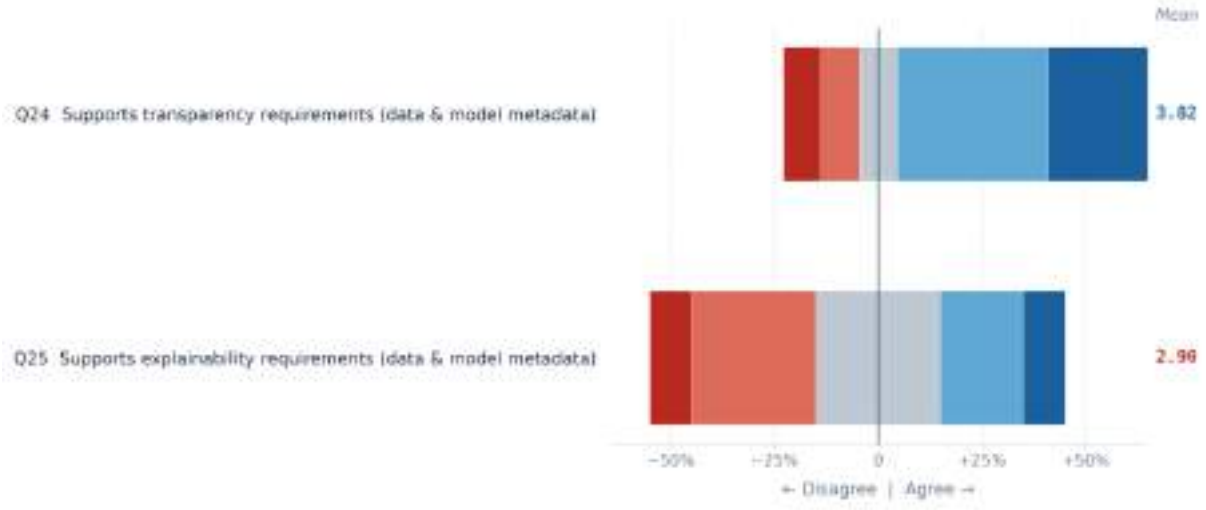


Figure 16 Response distributions and mean scores for AI Trustworthiness items (Q24–Q25).

In addition to the formal survey, informal feedback was gathered through bilateral LSP meetings, pilot debriefs, and emails. Key themes emerging from these interactions were consistent with the formal results, particularly regarding the need for asset duplication functionality and clearer phase naming. These inputs are incorporated into the actionable findings below.

3.2.2 Analysis: Key Actionable Findings

Finding 1 — Strong foundation in Usefulness and Trust. The two highest-performing dimensions were Perceived Usefulness (μ = 3.68) and Trust & Confidence (μ = 3.68). Users recognise the Hub as a valuable tool for transparency and documentation and trust its outputs. This provides a strong foundation for continued adoption.

Finding 2 — Explainability support is the critical gap. Q25 (μ = 2.90) is the only item below the neutral midpoint and the lowest score across the entire survey. The AI Model Hub supports transparency (recording what and how) but not yet explainability (documenting why a model behaves as it does). A related structural gap was identified for use cases based on pre-trained LLMs, where the model passport fields (which are designed primarily around training parameters) do not map well to pre-trained or fine-tuned model architectures. Adapting the passport schema for these cases is a complementary requirement to explainability.

Finding 3 — Technical accessibility limits reach to non-technical users. Q13 (μ = 3.23) and qualitative responses consistently highlight that risk assessors without a technical background struggle to interpret passport fields and understand their purpose. The interface is intuitive for developers but not yet self-explanatory for risk assessors.

Finding 4 — Weak social and organisational embedding. Q19 (μ = 3.09) and Q20 (μ = 3.27) indicate that the Hub has not yet achieved strong peer or management visibility within LSP organisations. This is expected at an early adoption stage but should be monitored.

Finding 5 — Workflow integration needs improvement. Q21 (μ = 3.25) and several open-text responses indicate friction in integrating the Hub into existing workflows, particularly around repetitive data entry for similar assets. One respondent also suggested the ability to share or compare asset data within and across organisations for benchmarking purposes, pointing to a longer-term requirement for collaborative data governance features.

Figure 17 provides a comparative overview of mean scores across all six TAM dimensions, illustrating where the AI Model Hub performs strongest and where gaps remain.

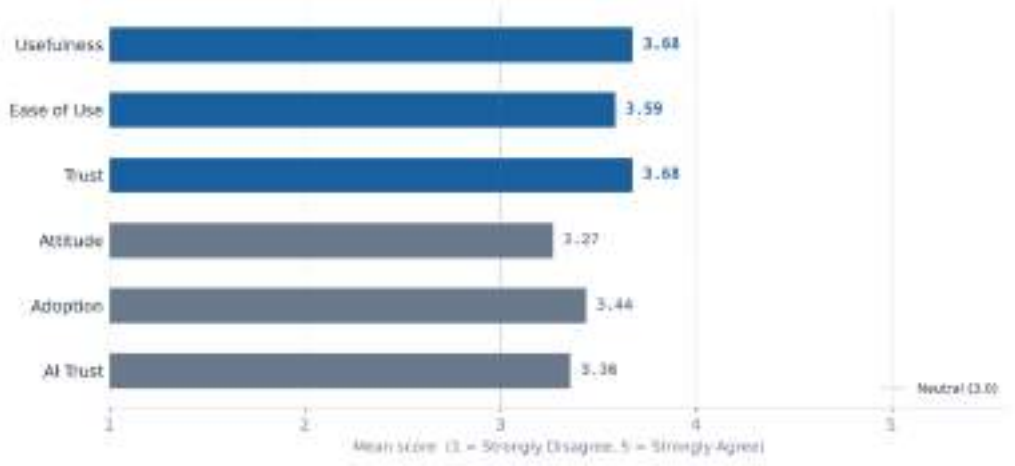


Figure 17 Mean scores per TAM dimension.

Figure 18 presents the same dimension scores in radar form, making cross-dimensional patterns immediately visible.



Figure 18 Radar plot overview of TAM dimension scores.

Table 8 Qualitative themes drawing from Q10/Q15/Q26.

Theme	Specific requests
Feature requests	Asset duplication, post-creation editing, asynchronous passport filling, Data Passport metadata schema extension. Cross-asset and cross-LSP data comparison for quality assurance and benchmarking purposes.
Integration	TrustGuard linkage, Model Passport to Risk Assessment linkage, pre-trained LLM support.
Guidance	In-context help text, risk assessor guidelines, phase naming clarification, verified vs. non-verified passport distinction (MLflow registration requirement).
Technical issues	Data passport entry errors (resolved), firewall access at SINTEF.

3.2.3 Updates Implemented & Planned

Based on the findings in Section 3.2.2, the following updates have been identified for the Version 2 release of the AI Model Hub, targeting the Replication Pilots:

Implemented prior to **Version 2**:

- *Clarified lifecycle phase naming to make it unambiguous that the production/deployment phase must appear last (addresses Q15 feedback)*
    - Clarification of the verified vs. non-verified AI Model Passport distinction delivered through targeted LSP training sessions. The verified AI model passports are automatically generated utilising the AIPassport pythonic library and the MLFLOW tracker during the development lifecycle of AI models. The latter implies that the lifecycle ends when a model is produced and registered on the MLFLOW tracker in order the process to generate automatically the AI model passport on the AI model Hub.

Planned for **Version 2**:

- *In-context guidance and tooltips across passport fields, specifically targeting non-technical risk assessors (addresses Finding 3, Q13, Q15 feedback)*
    - The responsible team already started to update the user manual of the AI Model Hub. The new manual will include more details and description about the information and metadata necessary to fill in on the AI Model Hub. The activity already started.
- *Asset duplication functionality allowing users to clone an existing model or dataset and modify only changed fields (addresses Finding 5, Q10 feedback)*
    - The responsible technical team already started working on the specific recommendation, aka the "duplication functionality". All technical teams are working to integrate a single sign-on Authentication and Authorization Infrastructure (AAI) to assist on the cross-tools services of FAITH digital ecosystem (AI model Hub, TrustGuard, SSM, TrustSense). The latter provides the capability to develop the Asset duplication functionality. The activity already started.
- *Post-creation editing of registered models and datasets (addresses Q10 feedback)*
    - While the recommendation addresses the feedback raised in Q10, it may not be fully aligned with the rationale of the AI Model Passports and the AI Model Hub, as unrestricted post-creation editing could compromise traceability, version control, and trust in registered models and datasets. The activity is planned after v2
- *Explainability support enhancements — extended metadata fields and/or structured sections to capture model behaviour explanations (addresses Finding 2, Q25)*
    - The responsible team already started to update the user manual of the AI Model Hub. The new manual will include more details and descritpion about the information and metadata necessary to fill in on the AI Model Hub. The activity already started.
- TrustGuard integration — linking risk assessments conducted in TrustGuard to Model Hub records (addresses Q10 feedback) // Model Passport — Risk Assessment linkage — dedicated section in the Model Passport for associated risk assessments (addresses Q10 feedback)
    - The activity is in advanced development phase..

3.3 TrustSense

3.3.1: Formal & Informal LSP Feedback Results

1. **Section 1: About You**

   The trustSense maturity assessment tool was validated across the 7 Large Scale Pilots (LSPs), capturing operational inputs from project partners, domain experts, and risk assessors representing sectors such as Wastewater Management (LSP5), Education (LSP3), Healthcare (LSP6), and Active Ageing (LSP7). Validation data was collected during the second workshop via a Full Evaluation Survey (n=4), a Quick Usability Feedback Survey (n=2), and live session trial logs in which all LSPs provided feedback.

2. **Section 2: Perceived Usefulness** (link with Technology Acceptance Model (TAM) 1)

   The quantitative response frequencies for Perceived Usefulness items are detailed below:

Table 9 Perceived Usefulness Response Frequencies

Survey Question	1	2	3	4	5	Mean
Q1: Understanding overall purpose/function	0%	25%	0%	25%	50%	4.00
Q2: Clarity of domains and definitions	0%	0%	50%	25%	25%	3.75
Q10: Interpretability of trustScore output	0%	0%	25%	50%	25%	4.00
Q11: Relevance of scoring and mitigations	0%	0%	50%	25%	25%	3.75

Note: Responses were recorded on a five-point Likert scale, where 1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, and 5 = Strongly agree.

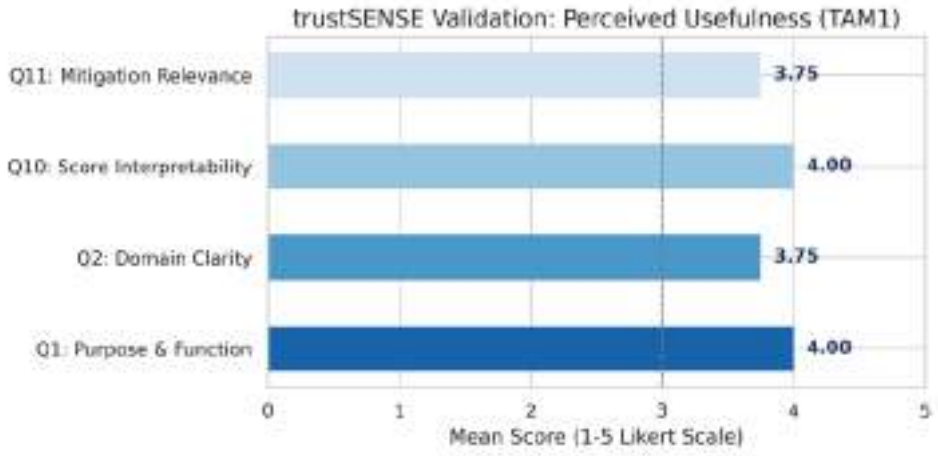


Figure 19 trustSense Perceived Usefulness.

3. **Section 3: Ease of Use** (link with TAM 1)

The quantitative response distributions for user interface and workflow ease of use are detailed below:

Table 10 Interface and Workflow Ease of Use Response Frequencies

Survey Question	1	2	3	4	5	Mean
Q12: Accessibility for non-technical teams	25%	50%	0%	25%	0%	2.25
Q16: Ease of use of UI and workflow	25%	25%	0%	25%	25%	3.00
Quick Survey Q1: Ease of following process	0%	0%	0%	50%	50%	4.50

The live run logs submitted directly by pilot risk assessors during the Workshop 2 live trials generated the following raw data parameters:

Table 11 Maturity level results

Participant	Chosen Profile	Maturity Level Result
LSP3	AI Technical Team	High
LSP6	AI Technical Team	High
LSP2	AI Technical Team	High
LSP1	Domain AI User Team	Moderate
LSP7	AI Technical Team	Moderate
UoS	Cybersecurity Defenders	Moderate
LSP4	Domain AI User Team	Moderate
LSP2	Domain AI User Team	High

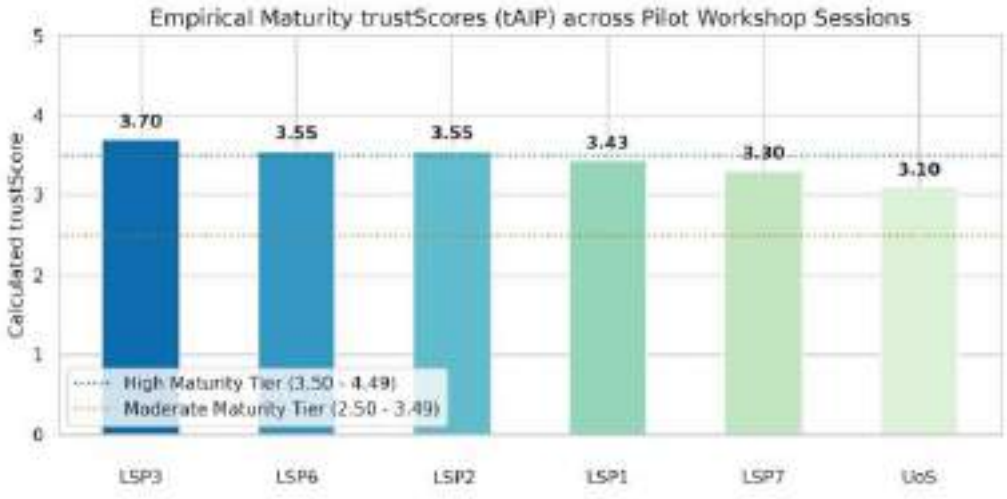


Figure 20 trustSense scores examples.

4. **Section 4: Human Factors** (link with TAMs 2 & 3)

The quantitative response frequencies for the Human Factors dimension (comprising adoption confidence and scoring weight validation) are itemized in the table below:

Table 12 Response Frequencies for the Human Factor dimension

Survey Question	1	2	3	4	5	Mean
Q14: Weight alignment with expert judgment	0%	0%	75%	25%	0%	3.25

Q17: Confidence in day-to-day adoption	25%	25%	0%	25%	25%	3.00

5. **Section 5: NIST Measures**

The trustSense instrument is designed to operationalise the NIST AI RMF GOVERN and MEASURE functions by evaluating human oversight capabilities (Accountability & Transparency) and translating collective team traits into an empirical maturity rating (tAIP).

During the validation trials, the tool's output was evaluated against its core risk-calibration framework defined in deliverable D2.1, where the human maturity rating functions as a direct modifier to baseline technical risks (R) to calculate the final calibrated risk posture (fR = T x V x I). The qualitative tracking from Workshop 2 monitored the operational readiness of these scores to serve as reliable risk inputs across varying lifecycle stages.

3.3.2 Analysis: Key Actionable Findings

Finding 1 — Operational Disconnect with Early-Lifecycle Practices

Because multiple pilot domains responded "NO" when asked if the questions reflected the realities of their daily activities, reporting that their use of AI is currently limited or not yet implemented, and that infrastructure management currently lacks formally instituted organisational procedures and teams at this stage, there is an identified baseline mismatch for early-lifecycle environments where specific operational metrics cannot yet be evaluated.

Finding 2 — Ambiguity From Arbitrary Temporal and Numerical Constraints

Because multiple expert reviewers explicitly flagged temporal and quantitative constraints as confusing (e.g., mandating a strict *"minimum of four training sessions annually"*, *"biweekly reviews"*, or *"90% challenge resolution ratios"*), a decision to strip specific numbers and replace them with the term "routine" was taken. This allows individual organisations to naturally evaluate adherence against their internal quality management standards rather than arbitrary windows.

Finding 3 — Confounding Items with Multiple Overlapping Themes

Because 25% of full-survey respondents flagged severe questionnaire ambiguities, noting that a single statement frequently mashed three distinct objectives together (such as grouping *operational business continuity* simultaneously with *project deadlines* and *90% success rates*), a decision to break apart all multi-theme items into single, isolated prompts was taken. This guarantees that a team fulfilling one metric but failing another can record an accurate score.

3.3.3 Updates Implemented & Planned

Consensus Workshop Protocol: Redesigned the trustSense onboarding documentation and user instructions to mandate a facilitated group-consensus methodology for small teams, eliminating unguided individual survey deployments.

Abstract Phrasing Refinement: Adjusted the vocabulary and contextual definitions of items within the *Responsibility* and *Governance* blocks based on feedback from project workshops to ensure clear, unambiguous interpretation across non-technical domain groups.

Standardised Layout Implementation: Refined the user interface to deliver a stable, standardised interface across all sections rather than an environment that alters in real time, ensuring a predictable user interaction path.

Methodological Guidance Expansion: Re-aligning the framework's documentation to provide clearer guidance on facilitation requirements, explicit operational assumptions regarding group-based scoring, and stronger integration of human and organisational context into the interpretation of results.

Harmonised Indicators Development: Planning the development of harmonized indicators within future qualitative research cycles to enable cross-tool comparisons and qualitative benchmarking once compatible organisational datasets become available.

3.4 TrustGuard

TrustGuard is an in-isolation AI risk assessment system designed to evaluate risks arising from the compromise of key trustworthiness characteristics—such as accuracy, robustness and security—across both model and data assets throughout the entire AI lifecycle, from design and development to deployment. In addition to identifying risks, the tool provides asset-level controls to support targeted mitigation.

Built as a checklist-based solution, TrustGuard operationalises a six-phase, risk-based methodology, aligned with ISO 27005 and ISO 42001 standards and leverages knowledge from the EU High-Level Expert Group on AI (HLEG), ENISA, NIST's AI Risk Management Framework, and relevant ISO/IEC initiatives.

3.4.1 Formal & Informal LSP Feedback Results

Section 1: About the Responders

The survey opened with demographic and contextual questions to characterise the respondent sample and provide context for interpreting the subsequent findings. A total of 13 responses were collected from participants representing the FAITH Large-Scale Pilots (LSPs), five partner organisations, and both developer and risk assessor roles. Most respondents evaluated AI systems that had already reached the deployment phase and were operating in real-world environments, ensuring that the feedback reflects practical experience with TrustGuard in operational settings.

Question	Key Findings
Q1 – LSP represented	Responses were obtained from six FAITH Large-Scale Pilots. The Media LSP was the most represented (5 responses; 38%), followed by the Transport LSP (3 responses; 23%) and the Health LSP (2 responses; 15%). The remaining three LSPs each contributed one response (8%), providing broad coverage across the project's pilot domains.
Q2 – Organisation represented	The 13 respondents represented five consortium organisations. Athens Technology Center (ATC) contributed the largest number of responses (5), while the remaining participants were distributed across ISTI-CNR, FORTH, UNIFI, UNIPI, ICCS, SINTEF and MERMEC Engineering.
Q3 – Primary role	The respondent pool consisted primarily of Developers (54%) and Risk Assessors (46%). This distribution indicates that the evaluation was predominantly performed by participants directly involved in AI system development and trustworthiness assessment.
Q4 – Lifecycle phase	The majority of respondents (77%; 10 of 13) evaluated AI systems in the Deployment Phase, while the remaining 23% (3 respondents) assessed systems still in the

	Development Phase, indicating that most evaluations were conducted on mature systems.
Q5 – Deployment environment	For respondents whose systems were already deployed, 89% (8 of 9) reported that their AI models were operating in real-world environments, while 11% (1 of 9) indicated deployment in a sandbox or testing environment. Four respondents did not answer this question because it was not applicable to systems still under development.

Table 13 Summary of demographic and contextual characteristics of survey respondents for TrustGuard (N=13). Counts marked with ~ are approximate, read from bar charts in the original survey report.

Figure 21 below illustrates the demographic profile of the survey respondents for TrustGuard.

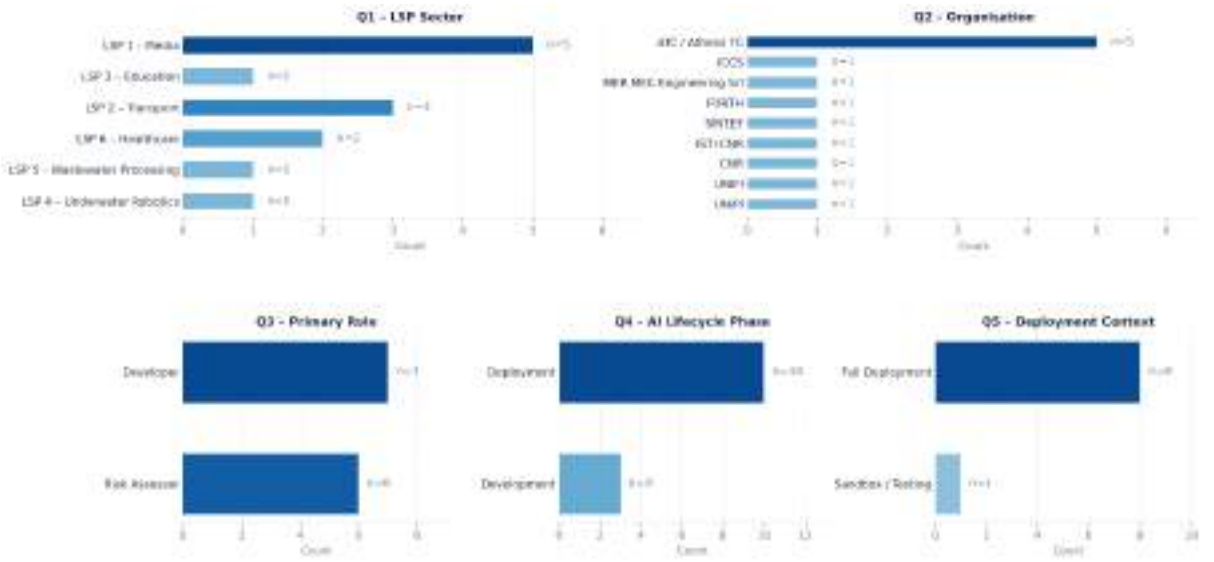


Figure 21 Demographic profile of the survey respondents for TrustGuard. Top row: distribution by LSP sector and organisation. Bottom row: primary role, AI lifecycle phase, and deployment environment.

Section 2: Perceived Usefulness (link with Technology Acceptance Model (TAM) 1)

Similarly to the other FAITH components, the Perceived Usefulness was assessed through four Likert-scale items (Q6–Q9), which examine whether TrustGuard improves the quality, productivity, efficiency, and the overall usefulness for identifying risk and assessing the trustworthiness of an AI asset.

The results, illustrated in Figure 22, indicate a positive perception of TrustGuard's usefulness, with an average score across the four items of μ = 4.17 / 5.0. The highest-rated item was Q7 (TrustGuard increases my productivity by making it easier to identify relevant threats and vulnerabilities than previous methods, μ = 4.33), with all respondents selecting either "Somewhat Agree" or "Strongly Agree". Q9 (Overall, I find TrustGuard useful for identifying risk and assessing the trustworthiness of an AI asset, μ = 4.22) and Q8 (TrustGuard enables me to complete tasks more efficiently than my usual methods because it provides a clear, structured set of questions, μ = 4.11) also received strong positive evaluations. Q6 (Using TrustGuard improves the quality of my in-isolation risk assessment work by helping me produce more complete and internally consistent risk assessment reports, μ = 4.00) obtained the lowest mean among the four items, although the majority of respondents still expressed agreement. Overall, the findings suggest that users perceive TrustGuard as a valuable tool that

enhances the quality, efficiency, and usefulness of the risk assessment process, while indicating some opportunity to further strengthen users' perceptions of its impact on assessment quality.

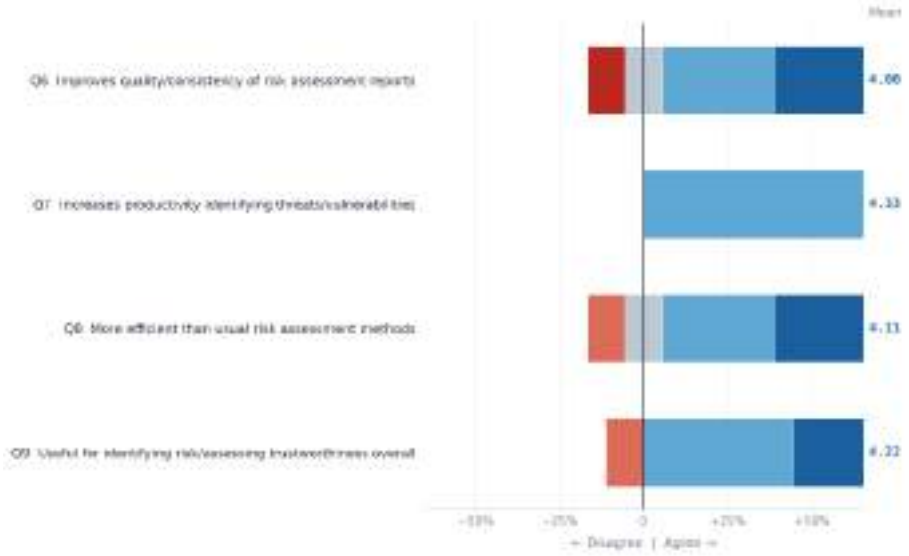


Figure 22 Response distributions for Perceived Usefulness items (Q6–Q9) for TrustGuard. Bars represent percentage of respondents per response category.

Section 3: Ease of Use (link with TAM 1)

Perceived Ease of Use was measured through four items (Q11–Q14) addressing learnability, intuitiveness, technical accessibility, and overall ease of use. The dimension mean was μ = 4.34 / 5.0, indicating a highly positive evaluation of TrustGuard's usability. The highest-rated item was Q11 (I found the TrustGuard easy to learn, μ = 4.78), suggesting that users considered the tool straightforward to become familiar with. Q14 (Overall, I find TrustGuard easy to use, μ = 4.56) and Q12 (Interacting with TrustGuard is intuitive, clear and understandable, μ = 4.44) also received strong agreement, demonstrating that respondents perceived the interface as intuitive and user-friendly. The comparatively lowest-rated item was Q13 (Using TrustGuard does not require high levels of technical knowledge or support, μ = 3.56), which exhibited the greatest variability in responses, including both disagreement and agreement. This indicates that, although users generally considered TrustGuard easy to learn and use, some respondents still perceived that effective use requires a certain level of technical expertise, which is an anticipated result, as TrustGuard as both vulnerability and threat assessment require highly technical knowledge (for instance the vulnerability assessment includes questions such as: “Has the model been configured to limit detailed output (e.g., probability scores, internal logic traces) that could reveal sensitive model characteristics or facilitate adversarial probing?”). This interpretation is further supported by the qualitative feedback, where participants suggested improvements such as providing brief explanations or definitions for threats through hover tooltips and tailoring questionnaires to different types of AI assets, which could further reduce the effort required to complete assessments.

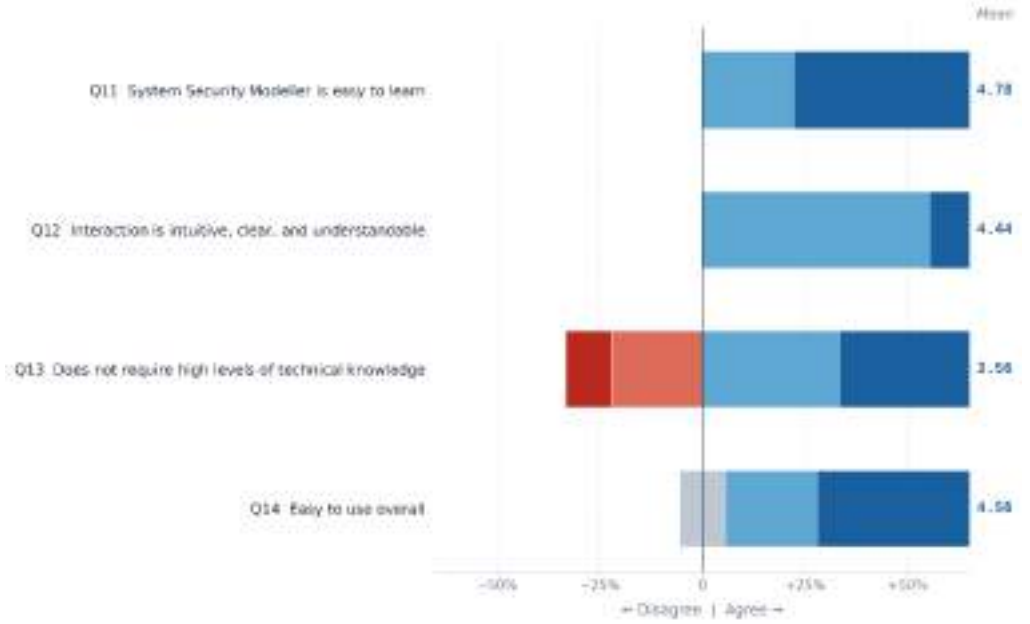


Figure 23 Response distributions for Easy of Use items (Q11–Q14) for TrustGuard. Bars represent percentage of respondents per response category.

Section 4: Human Factors (link with TAMs 2 & 3)

Trust & Confidence (Q17–Q18)

Respondents expressed a high level of confidence and trust in TrustGuard, with a sub-dimension mean of μ = 4.19 / 5.0. Participants reported feeling confident in their ability to use the tool correctly in their day-to-day work (Q17, μ = 4.12), with the majority selecting either "Somewhat Agree" (63%) or "Strongly Agree" (25%). Trust in the outputs and recommendations produced by TrustGuard was rated slightly higher (Q18, μ = 4.25), with 88% of respondents expressing agreement. Only a small proportion of respondents selected the neutral option for either question, indicating that users generally perceive the tool as both reliable and easy to use correctly.

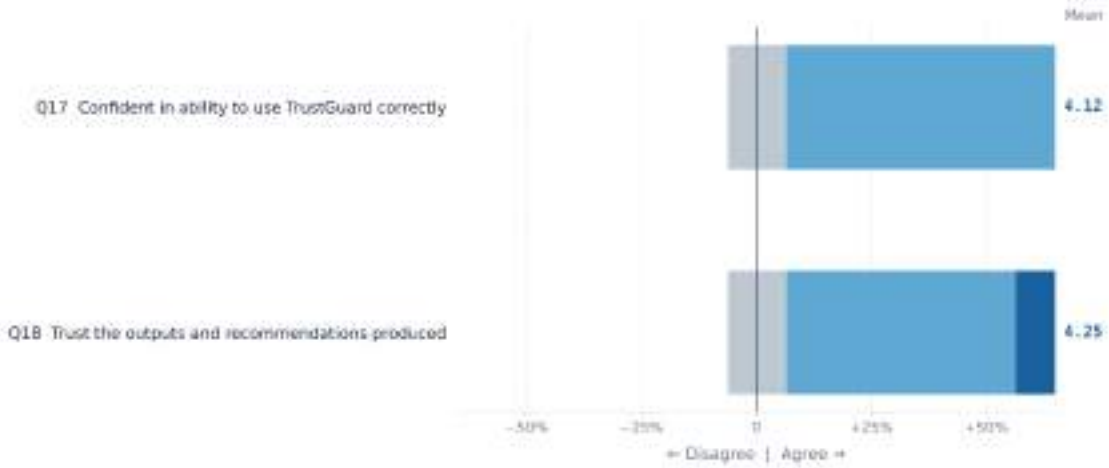


Figure 24 Response distributions for Easy of Use items Q17, Q18 for TrustGuard. Bars represent percentage of respondents per response category.

Attitude & Social Influence (Q19–Q21)

Overall attitudes towards TrustGuard were positive, with respondents indicating that they liked the tool (Q19, μ = 4.00). Perceptions of organisational and peer support were somewhat more moderate.

Respondents (76%) generally agreed that senior staff or management supported the use of TrustGuard (Q21, μ = 3.88), while perceptions of colleagues' views were slightly lower (Q20, μ = 3.62). The latter question exhibited the greatest spread of responses within this section, including some disagreement and neutral responses, suggesting that although individual acceptance of TrustGuard is high, awareness and recognition among peers may still be developing as deployment expands across the pilots.

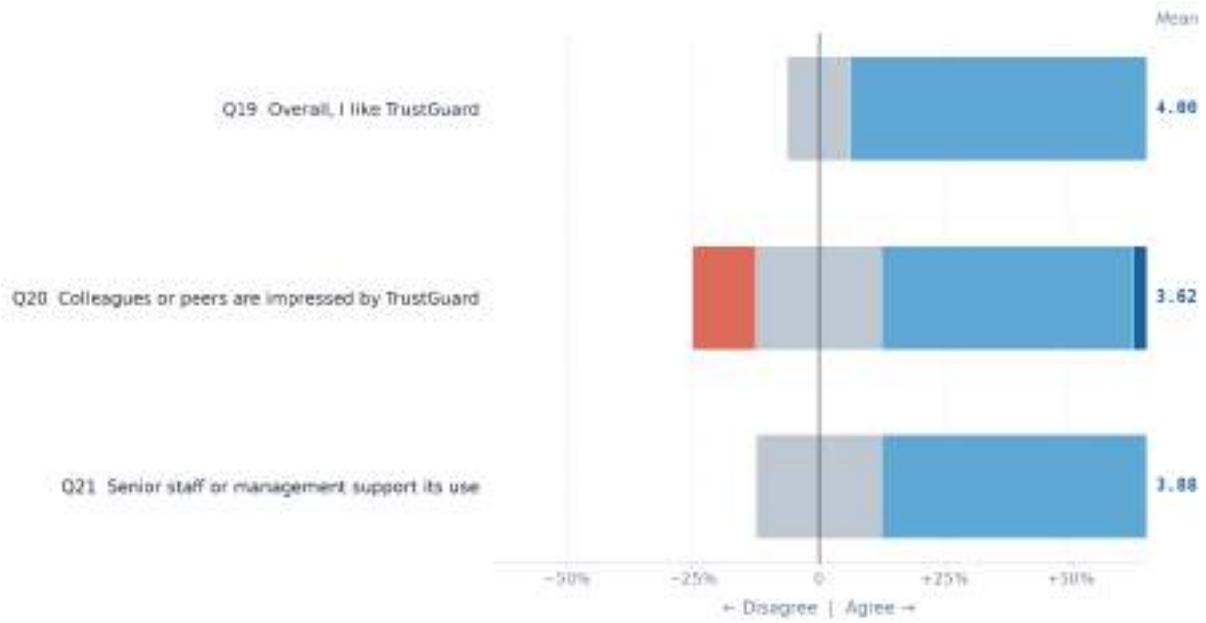


Figure 25 Response distributions for Easy of Use items (Q19–Q21) for TrustGuard. Bars represent percentage of respondents per response category.

Compatibility & Behavioural Intention (Q22–Q23)

Compatibility with existing workflows and future adoption intentions received positive but comparatively lower ratings than the previous dimensions, with a combined mean of μ = 3.63 / 5.0. TrustGuard's fit with existing workflows and organisational processes (Q22, μ = 3.50) was the lowest-rated item in the Human Factors section and showed the greatest variability in responses, including disagreement from some participants. This suggests that integration into established risk assessment practices remains an area for improvement. Nevertheless, respondents indicated a positive intention to continue using TrustGuard in future work (Q23, μ = 3.75), with 75% selecting either "Somewhat Agree" or "Strongly Agree". These findings suggest that users recognise the value of the tool and are willing to continue using it, although further efforts to streamline integration into existing organisational processes could facilitate broader adoption. It should be noted, however, that the current evaluation did not capture respondents' prior experience with conducting AI risk assessments. As a result, it is not possible to determine whether perceptions of workflow compatibility or ease of adoption were influenced by familiarity with risk assessment processes rather than by the tool itself. Including questions on participants' previous experience with AI risk assessment in future evaluations would enable a more nuanced interpretation of the results and help distinguish between challenges related to the methodology and those related to TrustGuard's implementation.



Figure 26 Response distributions for Easy of Use items (Q22–Q24) for TrustGuard. Bars represent percentage of respondents per response category.

Section 5: NIST Measures

Five items (Q24–Q28) evaluated the extent to which TrustGuard supports AI trustworthiness activities aligned with the NIST AI Risk Management Framework (AI RMF) – MEASURE function, as defined in D3.3. Specifically, the questions correspond to Measure 2.1 (Documentation), Measure 2.6 (AI system is evaluated regularly for safety risks), Measure 2.7 (AI system security and resilience are evaluated), Measure 2.8 (Risks associated with transparency are examined), and Measure 2.11 (Fairness and bias are evaluated).

Overall, respondents expressed positive perceptions across all five measures. The highest-rated items were Q25 (Supports regular and consistent risk assessments, μ = 16.00), corresponding to Measure 2.6, and Q27 (Supports the assessment of transparency risks, μ = 16.00), corresponding to Measure 2.8. For both questions, two-thirds of respondents selected either Somewhat Agree or Strongly Agree, indicating that TrustGuard is perceived as effectively supporting systematic AI risk assessment activities and the evaluation of transparency-related risks. These findings are consistent with TrustGuard's structured assessment methodology, which guides users through predefined threat, consequence and vulnerability assessments and produces documented risk assessment results.

Similarly, respondents reported positive perceptions regarding Measure 2.1 (Documentation). Q24 (Supports the documentation actions required during risk assessments, μ = 15.86) received predominantly positive responses, with 71% of participants selecting either Somewhat Agree or Strongly Agree. This suggests that users consider TrustGuard effective in supporting the documentation required during AI trustworthiness assessments, reflecting one of the primary objectives of the platform to provide a structured and traceable risk assessment process.

The remaining two measures also received favourable evaluations. Q26 (Supports evaluation of AI system resilience), corresponding to Measure 2.7, and Q28 (Supports assessment of bias in the AI system), corresponding to Measure 2.11, both achieved positive overall ratings. Nevertheless, these two questions exhibited a larger proportion of Neither agree nor disagree (or N/A) responses than the previous measures. This may indicate that some respondents had limited opportunities to evaluate these aspects within their specific pilot use cases or that resilience and bias assessments were not equally relevant across all AI systems evaluated.

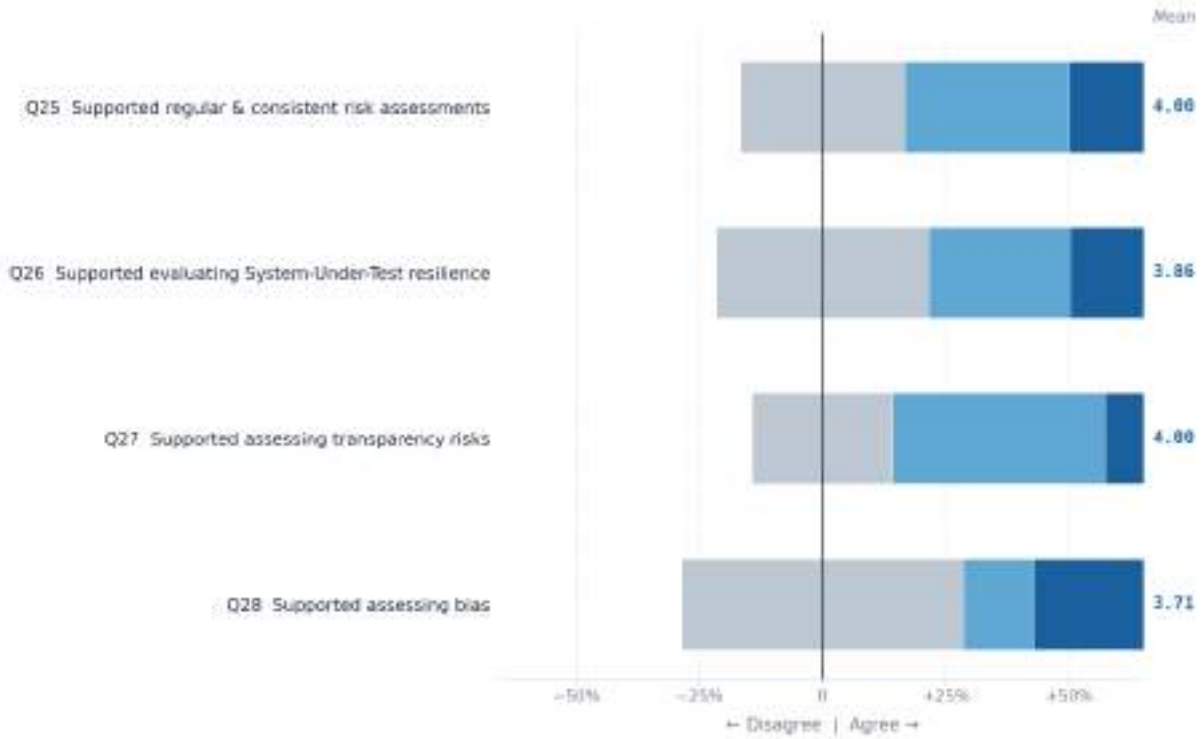


Figure 27 Response distributions for Easy of Use items (Q25–Q28) for TrustGuard. Bars represent percentage of respondents per response category.

Overall, the results indicate that respondents perceive TrustGuard as providing meaningful support for the NIST AI RMF MEASURE function. In particular, users recognised its contribution to structured documentation, consistent risk assessment practices, and the evaluation of transparency-related risks. While support for resilience and bias assessment was also viewed positively, the higher proportion of neutral or N/A responses suggests that future evaluations should further investigate how these capabilities are applied across different pilot domains and AI system types. This could be achieved by collecting additional contextual information on the types of assessments performed and the relevance of each trustworthiness dimension to the evaluated AI systems.

3.4.2 Analysis: Key Actionable Findings

Finding 1 — Strong overall user acceptance and trust.

The highest-performing dimensions were Perceived Ease of Use (μ = 4.34), Perceived Usefulness (μ = 4.17), and Trust & Confidence (μ = 4.19). Respondents consistently agreed that TrustGuard is easy to learn, intuitive to use, and supports the identification of AI trustworthiness risks more effectively than previous methods. High levels of trust in the tool's outputs and confidence in using it indicate a strong foundation for continued adoption across the FAITH pilots.

Finding 2 — Workflow integration remains the main adoption challenge.

While respondents expressed a clear intention to continue using TrustGuard (Q23, μ = 3.75), the lowest score within the Human Factors section was obtained for compatibility with existing workflows (Q22, μ = 3.50). This suggests that users recognise the value of the tool but still encounter challenges integrating it into established organisational processes. Future development should therefore prioritise improving interoperability with existing risk assessment practices and reducing friction during routine use.

Finding 3 — Technical accessibility can be further improved.

Although TrustGuard was rated highly for learnability (Q11, μ = 4.78), intuitiveness (Q12, μ = 4.44), and overall ease of use (Q14, μ = 4.56), respondents were less positive regarding the level of technical knowledge required (Q13, μ = 3.56). Open-text feedback also highlighted the need for contextual guidance, clearer definitions of threats and vulnerabilities, and additional explanations through hover tooltips. These findings suggest that targeted usability enhancements could further improve accessibility, particularly for users with less technical expertise.

Finding 4 — TrustGuard effectively supports the NIST AI RMF MEASURE function.

The highest ratings within the NIST Measures section were obtained for Measure 2.6 (Regular and consistent risk assessment, μ = 16.00) and Measure 2.8 (Transparency risk assessment, μ = 16.00), while documentation support (Measure 2.1, μ = 15.86) and resilience evaluation (Measure 2.7, μ = 15.86) were also positively evaluated. These findings indicate that respondents perceive TrustGuard as providing meaningful support for structured AI trustworthiness assessments and the documentation required by the NIST AI RMF.

Finding 5 — Future evaluations should better capture user background and deployment context.

Some survey items—particularly those related to workflow compatibility, resilience, and bias assessment—received a relatively higher proportion of neutral or N/A responses. Since the questionnaire did not collect information about respondents' previous experience with AI risk assessment, it is not possible to determine whether these responses reflect limitations of TrustGuard or differences in participants' familiarity with structured risk assessment methodologies. Future evaluations should therefore include questions on prior AI risk assessment experience, the frequency with which respondents conduct such assessments, and the relevance of individual trustworthiness dimensions to their specific AI systems. This would enable a more nuanced interpretation of user perceptions and facilitate comparisons across different stakeholder groups and pilot domains.

3.4.3 Updates Implemented & Planned

Based on the findings presented in Section 3.2.2, several improvements have been identified for the next release of TrustGuard, supporting its use during the in-context risk assessment activities of the Replication Pilots. The planned updates address both the user feedback collected through the evaluation and the recommendations gathered from pilot partners and WP2 discussions.

Updates Implemented prior to Version 2

- ***Improved wording of threat assessment questions and response scales (addresses Q10, Q15)***: The wording of several threat assessment questions has been revised by replacing domain-specific frequency scales with more generic likelihood-based wording where appropriate. Minor typographical issues identified during the evaluation have also been corrected.
- ***Enhanced Risk Assessment user interface (addresses Finding 2)***. The Risk Assessment interface has been redesigned according to the updated mock-ups presented and discussed in WP2 meetings, to improve navigation and better support users throughout the assessment workflow. Planned improvements include presenting identified risks in descending order of criticality, thereby helping users prioritise mitigation activities more efficiently and avoiding information repetition. Additional information will be provided to explain how overall risk scores are calculated based on the assessed threat likelihood, consequence severity and vulnerability levels (addresses Q29). This enhancement aims to improve users' understanding of the assessment methodology and further strengthen trust in the generated results.

- ***Improved calculation of risk-assessment.*** The handling of non-applicable controls has been improved so that they are treated appropriately within the vulnerability assessment and final risk calculations (addresses Q29).
- ***Improved wording of consequence assessment questions to align with FAITH TAF v2 and facilitate integration with the System Security Modeller (SSM)***. The consequence assessment questionnaire has been revised to align with the updated FAITH Trustworthiness Assessment Framework (FAITH_TAF_v2). The revised wording provides greater consistency across the trustworthiness assessment methodology while also harmonising the consequence categories and terminology with those adopted by the System Security Modeller (SSM). This alignment simplifies the exchange of assessment results between the two tools and supports the integrated FAITH risk assessment workflow.

Planned for Version 2

- ***Revision of threat and vulnerability assessment questions (addresses Q10, Q15)***: The questions will be updated following the developments of the FAITH ontology.
- ***Contextual guidance and threat explanations (addresses Finding 3, Q10, Q13, Q15).*** Brief explanations and definitions for threats will be provided directly within the user interface (e.g., through hover tooltips or contextual help). This recommendation originated both from the user evaluation and from pilot feedback and aims to improve accessibility for users with limited experience in AI risk assessment. In the longer term, these explanations will be aligned with the FAITH ontology currently under development.
- ***Questionnaires tailored to AI system characteristics (addresses Q10, Q16).*** The threat and vulnerability assessment questionnaires will be progressively adapted according to the type of AI asset, AI model, and lifecycle phase. This will reduce the number of irrelevant questions presented to users and improve the applicability of TrustGuard across different pilot domains. The implementation is planned following completion of the FAITH ontology, as it will based on the knowledge formally represented in the ontology.
- ***Improved management of repeated risk assessments (addresses Finding 2, Q10).*** Several features are planned to facilitate iterative risk assessments, including support for copying existing assessments, preserving previous responses where appropriate and freezing completed assessments to preserve traceability while allowing users to create updated assessment versions.
- ***Improved handling of mitigation actions and vulnerabilities (addresses Q10).*** The relationship between identified vulnerabilities and the corresponding mitigation actions will be further improved to improve consistency and reduce potential confusion during the interpretation of assessment results. This implementation is planned following completion of the FAITH ontology, as it will based on the knowledge formally represented in the ontology.
- ***Integration with the FAITH Knowledge Base and TrustSense.*** Future versions of TrustGuard will benefit from updates to the FAITH ontology, including ontology-driven threat definitions and improved filtering of assessment questions.

3.5 FAITH Tools Validation Summary

The three FAITH in-isolation trustworthiness assessment tools, AI Model Hub and Passport, TrustSense, and TrustGuard, were deployed during the Phase 1 Pilots across all seven LSPs and validated through formal TAM- and NIST-aligned surveys, supplemented by informal feedback gathered through training sessions, WP2 meetings, and direct partner exchanges. Taken together, the results point to a consistent picture: strong underlying trust and perceived usefulness across all three

tools, with accessibility for non-technical users and integration into existing workflows as the two recurring, addressable barriers to adoption.

AI Model Hub and Passport received 13 valid formal responses across all seven LSPs. Perceived Usefulness and Trust & Confidence were its highest-rated dimensions (μ = 3.68/5 each), while explainability support was the only item to fall below the neutral midpoint (Q25, μ = 2.90/5) — a gap with direct relevance to NIST and EU AI Act compliance requirements, and the tool's clearest development priority. Technical accessibility for non-technical risk assessors (Q13) and workflow integration, particularly repetitive data entry (Q21), were the other main friction points.

TrustGuard showed the strongest overall acceptance of the three tools, with Perceived Ease of Use (μ = 4.34/5), Trust & Confidence (μ = 4.19/5), and Perceived Usefulness (μ = 4.17/5) all rated highly, and strong support recorded for the NIST AI RMF MEASURE function. As with the AI Model Hub, workflow compatibility (Q22, μ = 3.50/5) was its lowest-scoring dimension despite a clear stated intention among respondents to keep using the tool (Q23, μ = 3.75/5), and technical knowledge requirements (Q13) were flagged as a secondary accessibility barrier.

TrustSense was validated through a distinct, more qualitative methodology (facilitated consensus-building workshops, simulated evaluation runs, controlled group deployments, and iterative digital tracking) reflecting its role as a team-level, consensus-based maturity assessment rather than an individually-scored tool. Its key findings were methodological rather than quantitative: a baseline mismatch for LSPs at an early stage of AI adoption where the tool's assumptions about formalised teams and procedures did not yet apply; ambiguity introduced by arbitrary numerical and temporal thresholds (e.g., "four training sessions annually"); and survey items that conflated multiple distinct themes into a single question.

Three patterns recur across all three tools and are worth flagging as consortium-wide priorities rather than single-tool fixes. First, accessibility for non-technical and non-specialist users is the most consistent barrier identified; the AI Model Hub's Finding 3, TrustGuard's Finding 3, and TrustSense's wording ambiguities all point to the same underlying need for clearer, in-context guidance and simplified language. Second, workflow and process integration is the main practical adoption barrier for both the AI Model Hub (Finding 5) and TrustGuard (Finding 2), both flagging friction with existing organisational practices rather than any doubt about the tools' underlying value. Third, all three tools are converging on deeper cross-tool integration as the primary direction for Version 2: a shared single sign-on authentication infrastructure across the AI Model Hub, TrustGuard, SSM, and TrustSense; TrustGuard's consequence-assessment wording has already been aligned with FAITH TAF-v2 and harmonised with the SSM; and both TrustGuard and the AI Model Hub have planned features (ontology-driven threat definitions, Model Passport–Risk Assessment linkage) that depend on the FAITH ontology currently under development. TrustSense's early-lifecycle mismatch finding also echoes a theme raised independently in the TAF-v2 validation (Section 2.2.2): LSPs not yet operationally using AI consistently find trustworthiness-assessment concepts more abstract, regardless of which FAITH tool is involved.

All three tools have already implemented near-term fixes arising from this feedback (clarified phase naming and passport-verification guidance for the Model Hub; revised threat-assessment wording and an improved risk-calculation logic for TrustGuard; a redesigned consensus-workshop protocol and simplified, de-conflated question wording for TrustSense) and have Version 2 roadmaps that converge on three shared priorities: in-context guidance and tooltips for non-technical users, tighter cross-tool integration via the shared authentication infrastructure and the FAITH ontology, and better support for iterative, repeated assessments (asset duplication, assessment versioning) to reduce workflow friction.

4. Conclusions and Next Steps

4.1 Overall validation conclusions

The validation activities for D3.4 confirm that both the FAITH Trustworthiness Assessment Framework (TAF-v2) and the suite of tools used so far for in-isolation assessment (i.e., AI Model Hub, TrustSense, and TrustGuard) have achieved a consistent consensus of being "*validated with minor refinements*".

The validation process, conducted across multiple stakeholder groups from the LSPs, demonstrates a strong foundation of credibility and usefulness. TAF-v2 was appreciated for its conceptual clarity, completeness, and internal coherence. Similarly, the assessment tools consistently showed high scores in perceived usefulness, ease of use, and trust, particularly in their role as enablers for transparency, documentation, and systematic risk assessment.

While the validation was highly positive, three cross-cutting priorities were identified for refinement before wider rollout:

1. **accessibility for non-specialists:** A recurring need for clearer in-context guidance, simplified language, and intuitive tooltips was identified across all tools and the framework itself.
2. **workflow integration:** Users identified friction in integrating these tools into existing organisational processes, highlighting a need for "light-touch" implementation and features such as asset duplication to reduce repetitive data entry.
3. **cross-tool integration:** The consortium-wide priority for Version 2 is to deepen integration through a shared authentication infrastructure, alignment with the FAITH ontology, and enhanced interoperability between tools (e.g., TrustGuard–SSM linkage).

The "*minor refinements*" rating reflects the fact that identified weaknesses are operational and addressable through documentation, usability improvements, and feature updates, rather than fundamental redesigns.

4.2 Future actions

The validation activities have provided a clear roadmap for the next phase of the FAITH project. To address the identified findings and operationalize the framework and tooling for wider rollout of the next version, the consortium will prioritize the following actions:

A. *Framework and documentation enhancement*

The focus will be on:

- developing an explicit narrative connecting each TAF-v2 phase to ensure intuitive flow
- producing a comprehensive glossary and simplified, non-specialist explanatory text for all framework phases
- creating worked, domain-specific examples, including quantitative calculations, to serve as practical references for LSPs.
- developing dedicated documentation and checklists for system-level, cascading-threat assessments.

B. *TrustGuard Evolution and Integration*

The next actions will incorporate the updates already implemented, mainly focused on improving usability and technical accuracy: (i) assessment questions were reworded for clarity and consistency with TAF-v2 and the SSM; (ii) the Risk Assessment interface was redesigned to prioritize critical risks

and improve navigation; and (iii) calculation logic was refined to accurately handle non-applicable controls.

Moreover, based on the evaluation feedback and WP2 recommendations, the following improvements are prioritized:

- leveraging the FAITH ontology to provide in-app contextual guidance, hover-tooltips, and automated filtering of assessment questions
- tailoring assessments of specific AI asset types and lifecycle phases to eliminate irrelevant questions and reduce user burden
- introducing features to copy assessments, preserve previous responses, and "freeze" completed versions to maintain traceability
- refining the relationship between vulnerabilities and mitigation actions to reduce interpretation errors
- strengthening integration with the FAITH Knowledge Base and TrustSense for unified assessment data

A subsequent phase of validation is planned, specifically designed to verify the efficacy of the refinements, to assess trustworthiness in more operational deployment, to measure impact and acceptance and to continuously capture feedback for continuous improvement (as per Tier 4 illustrated in Figure 1). Particularly, the validation effort will feature core activities including: (i) conducting an inter-rater consistency check to ensure the reproducibility and reliability of TAF-v2 assessments; (ii) evaluating the updated SSM tool in real-world scenarios to confirm its capacity for in-context, whole-system assessment; (iii) organising a final validation workshop. This workshop will engage a diverse, large-scale group of stakeholders from the LSPs, providing a dedicated forum for collective review, stress-testing of the refined framework, and final consensus-building on the operational readiness of the full FAITH tool-chain.

4.3 Link to subsequent deliverables and support to sustainability and exploitation activities

The results and activities documented in this report serve as a critical foundation for the finalization of the FAITH framework and its subsequent transition into the exploitation phase. By validating the TAF-v2 and its associated tool-chain (AI Model Hub, TrustSense, TrustGuard, and the forthcoming SSM), we have established a robust, evidence-based methodology that is ready for broader adoption.

The work presented here directly informs and supports the upcoming deliverables, particularly D3.5 “Validation Report v3” and the concluding reports from the LSPs.

The roadmap planned in D3.3 and the future activities detailed in this report provide the operational baseline for D3.5. The planned inter-rater consistency checks and the outcomes from the final LSP workshop will constitute the primary empirical data for the final evaluation of the FAITH framework, providing the necessary assurance that the solution is mature, reproducible, and ready for deployment. The findings of this report demonstrate the validity and utility of the TAF framework, confirming its suitability for diverse AI deployment environments. As the framework will be made publicly available, this validation work acts as the key proof-of-concept required to attract external users and maintain long-term ecosystem interest.

Beyond technical validation, this report captures the essential co-creation work undertaken with the LSPs. This "feedback loop" methodology, where user needs directly shape framework evolution, serves

as the foundation for the long-term sustainability of the FAITH tool-chain, ensuring that future updates remain aligned with practical, real-world requirements.

The iterative improvements (such as the refinement of assessment logic, interface redesigns, and the introduction of adaptive questionnaire) will serve for reducing the operational burden for users. This ensures that the LSPs can transition from pilot-testing to sustainable use, utilizing the tool-chain as a trusted asset for their internal risk management and compliance processes.

By systematically addressing identified gaps and planning for future usability enhancements, the FAITH consortium is ensuring that the framework is not merely a static output, but a dynamic, exploitable asset capable of evolving alongside the regulatory and technological landscape.

References

Arp, R., Smith, B., & Spear, A. D. (2015) *Building Ontologies with Basic Formal Ontology*. MIT Press, Cambridge, Massachusetts, USA.

Jackson, R. C., Balhoff, J. P., Douglass, E., Harris, N. L., Mungall, C. J., & Overton, J. A. (2019) ROBOT: A Tool for Automating Ontology Workflows. *BMC Bioinformatics*, 20(Article 407), 1–10. [https://doi.org/10.1186/s12859-019-3002-3]

Jensen, M., De Colle, G., Kindya, S., More, C., Cox, A. P., & Beverley, J. (2024) *The Common Core Ontologies*. 14th International Conference of Formal Ontology in Information Systems, Enschede, The Netherlands. [https://doi.org/10.3233/FAIA241292]

Otte, J. N., Beverley, J., & Ruttenberg, A. (2022) BFO: Basic Formal Ontology. *Applied Ontology*, 17(1), 17–43. [https://doi.org/10.3233/AO-220262]