Ref. Ares(2026)4470184 - 30/04/2026

D7.2

Report on the initial pilot - underwater robotics

Related Work Package	WP7 – Implementation of drones LSP
Related Task	T7.2 – Initial Pilot
Lead Beneficiary	APRA
Contributing Beneficiaries	UNIFI, VPF, APRA
Document version	V1.0
Deliverable Type	R
Distribution level	PU
Contractual Date of Delivery	M28
Actual Date of Delivery	

Authors	Alberto Topini, Alessandro Ridolfi, Andrea Minardi, Mark Tanner, Daniel Diosdado
Contributors	
Reviewers	Evangelia Anagnostopoulou (ICCS), Dimitris Koulentianos (EA)

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
0.1	First draft submitted for internal review process	24/04/2026
1.0	Final draft after internal review process	30/04/2026

Statement of Originality

This deliverable contains original unpublished work except where clearly indicated otherwise.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

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 reports the results of the initial pilot of the FAITH underwater robotics Large-Scale Pilot (LSP) 4, implemented in the Port of Ravenna to validate the first deployment of trustworthy AI-enabled Autonomous Underwater Vehicles in realistic port conditions. The pilot addressed two complementary AI applications developed by University of Florence (UNIFI): Automatic Target Recognition (ATR) for sonar-based underwater target detection and Autonomous Coverage (AC) for autonomous survey planning and bathymetric mapping. The objective was to assess both technical performance and trustworthiness in a safety-relevant maritime environment.

The pilot was carried out through a staged sequence of experimental campaigns, progressing from environmental contextualisation and dataset collection to testing with trained models and direct interaction with operational stakeholders, notably the Port Authority of Ravenna and the Ravenna Fire Brigade. This progressive approach enabled alignment between the original pilot design and the actual environmental, logistical and operational constraints of the port domain, while also supporting stakeholder awareness, user readiness and adoption of a human-on-the-loop operational model.

From a technical perspective, the pilot produced encouraging results. For ATR, dedicated high-frequency and low-frequency Forward-Looking Sonar (FLS) datasets were collected, annotated and used to train complementary instance-segmentation models. The high-frequency model achieved strong performance on unseen pilot data, with 95.6% precision, 97.6% recall, 96.0% F1-score and 98.2% mAP@0.5, while maintaining near-real-time onboard execution. The low-frequency model showed lower but still solid performance and extended the effective observation range up to approximately 10 m, supporting a dual-sonar workflow for wider-area search and close-range confirmation. For AC, the planner reached the requested target coverage in 100% of the tested validation areas and demonstrated a clear reduction in mission duration compared with an equivalent manned procedure, although the current mission success rate remains constrained by the prototypal maturity of the vehicle platform.

A major contribution of the pilot was the first structured application of the FAITH trustworthiness toolchain to the underwater robotics use case. TrustGuard, trustSENSE, and the FAITH AI Model Hub were used to assess AI and human assets, document the ATR high-frequency dataset and model through the Data Passport and AI Model Passport, and identify the main residual risk areas, particularly around Validity and Reliability, Safety, Explainability and Interpretability, and Accountability and Transparency. The UNIFI AI

Technical Team was assessed by trustSENSE at a High trust level, confirming a strong organisational basis for the next improvement cycle.

Overall, the initial pilot provides a credible baseline for the underwater robotics branch of FAITH. ATR already shows high maturity in sonar-based underwater perception, while AC demonstrates promising capability for autonomous bathymetric survey execution. At the same time, the pilot confirms the need for further work on system robustness, platform stabilisation, broader field validation and replication in a new operational context. These results directly support the preparation of the replication pilot in the Port of Valencia.

Table of Contents

1. Introduction 11
1.1 Scope and objectives of the 1st pilot in the Implementation of Drones LSP 11
1.2. Structure of the document 12
1.3. Evolution of the AI System Under Test (SUT) from initial setup 13
2. Progress on the AI system deployment(s) and execution of the Pilots 19
2.1. Status of the 1st Pilot execution 19
2.2. Status of the Replication Pilot preparation/initialization 38
2.3. Deviations from the original deployment plan 38
2.4. Data collection throughout the 1st Pilot execution 40
2.5. Integration into LSP workflows: Lessons learned 46
3. Application of FAITH Tools 48
3.1. Application of the FAITH AI TAF V1 methodology 48
3.2 Tool Validation Planning & Training Processes 56
3.3 FAITH Tool Validation Activities 75
3.4. Threat Analysis 87
3.5. Initial Risk Assessment results 89
3.6. Countermeasures 95
3.7 Formal Tool Feedback 97
4. AI System Evaluation 99
4.1. Quantitative metrics on pilot data 99
4.2. Trustworthiness Assessment 111
4.3 Preliminary Impact Evaluation 118
5. Regulatory and Ethical Assessment 121
5.1. Regulatory Alignment with the EU AI Act and Other Relevant Frameworks 121
5.2. Port-specific and Local Compliance 123
5.3. Technical Compliance of Data and AI Models 125
6. Conclusions and Roadmap for the Final Pilot Phase 126
6.1. Summary of achievements in the second period 126

6.2 Plan for Replication Phase.....128
6.3. Identified risks for the Replication Phase 129
7. References 131

List of Figures

Figure 1: Example of the evolved ATR AI SUT over an experimental campaign 15
Figure 2: Example of the Explainability report per the evolved ATR AI SUT (Part 1).....16
Figure 3: Example of the Explainability report per the evolved ATR AI SUT (Part 2).....16
Figure 4: Example of the Explainability report per the evolved ATR AI SUT (Part 3).....17
Figure 5: Example of the evolved AC AI SUT over an experimental campaign.....18
Figure 6: Map of the Port of Ravenna.....20
Figure 7: Port of Ravenna area.....20
Figure 8: "Southern breakwater" test location.....21
Figure 9: "Firefighters Port Base" test location.....22
Figure 10: Visual evidences of the first experimental campaign 23
Figure 11: Illustrations of the second experimental campaign.....24
Figure 12: Visual reports of the third experimental campaign.....25
Figure 13: Visual reports from the fourth experimental campaigns.....26
Figure 14: Visual evidences of ATR SUT deployment on low-frequency FLS frames over the third and fourth experimental campaign.....27
Figure 15: Visual evidences of ATR SUT deployment on high-frequency FLS frames over the third and fourth experimental campaign.....27
Figure 16: Visual evidences of AC SUT deployment on validation areas over the third and fourth experimental campaign.....28
Figure 17: Visual reports from the first workshop.....30
Figure 18: Visual example of on-field training activity 32
Figure 19: Visual reports of the second workshop/in-depth training session.....33
Figure 20: Example of tire target placed on the seabed. On the left, the high-frequency frame; on the right, the low-frequency one.....43
Figure 21: Example of mannequin target placed on the seabed. On the left, the high-frequency frame; on the right, the low-frequency one.....44
Figure 22: Example of sewer target placed on the seabed. On the left, the high-frequency frame; on the right, the low-frequency one.....45
Figure 23: ATR Dry-run cartography 57
Figure 24: ATR Dry-run criticality assessment 58
Figure 25: ATR Dry-run Frequency of threats 59
Figure 26: ATR Dry-run Business Consequence Level 60
Figure 27: ATR Dry-run Control Questions.....61
Figure 28: AC Dry-run Cartography 62
Figure 29: AC Dry-run Criticality Level.....63

Figure 30: AC Dry-run Frequency of Threats 65
Figure 31: AC Dry-run Business Consequence Level 66
Figure 32: AC Dry-run Control Questions Design 67
Figure 33: Example of the developed plan 69
Figure 34: Example of developed plan 70
Figure 35: UNIFI trustSENSE results 78
Figure 36: UNIFI trustSENSE results 79
Figure 37: Snippet example from the Data Passport 81
Figure 38: Snippet example from the AI Model Passport 82
Figure 39: Visual snippet from TrustGuard with the two reported in-isolation assessments 83
Figure 40: Example of TrustGuard Criticality Assessment 84
Figure 41: Example of TrustGuard Threat Assessment Questionnaire 85
Figure 42: Example of TrustGuard Consequence Assessment Questionnaire 86
Figure 43: Example of TrustGuard Vulnerability Assessment Questionnaire 87
Figure 44: TrustGuard results for the FAITH WP7 FLS Instance Segmentation Study 92
Figure 45: TrustGuard results for the FAITH WP7 FLS Instance Segmentation Study 92
Figure 46: TrustGuard results for High Frequency FLS Seabed Inspection Training Dataset 94

List of Tables

Table 1: Location 1 features 21
Table 2: Location 2 features 22
Table 3: Performance of the high-frequency FLS ATR model on the unseen test split derived from the first three experimental campaigns.....100
Table 4: ATR high-frequency FLS model: robustness check on the last experimental campaign with differently positioned targets 102
Table 5: ATR low-frequency FLS model: evaluation on unseen test split derived from the first three experimental campaigns.....104
Table 6: ATR low-frequency FLS model: robustness check on the last experimental campaign with differently positioned targets 105
Table 7: AC validation areas used for field testing.....107
Table 8: AC quantitative metrics on pilot data 110

List of Abbreviations

Abbreviation	Explanation
FAITH	Fostering Artificial Intelligence Trust for Humans
AC	Autonomous Coverage
AI	Artificial Intelligence
ATR	Automatic Target Recognition
AUV	Autonomous Underwater Vehicle
CNN	Convolutional Neural Network
CPP	Coverage Path-Planning
FLS	Forward-Looking Sonar
FNR	False Negative Rate
FPR	False Positive Rate
GDPR	General Data Protection Regulation
GUI	Graphical User Interface
IoU	Intersection over Union
KPI	Key Performance Indicator
LSP	Large Scale Pilot
mAP	Mean Average Precision
ROS	Robot Operating System
ROV	Remotely Operated Vehicle
SBP	Sub Bottom Profilers
SSS	Side-Scan Sonar
SUT	System Under Test
TRL	Technology Readiness Level

1. Introduction

1.1 Scope and objectives of the 1st pilot in the Implementation of Drones LSP

The initial pilot of the FAITH underwater robotics Large-Scale Pilot is aimed at validating the initial deployment of trustworthy AI-enabled AUVs in port environments, with a particular focus on the Port of Ravenna. The pilot is positioned within WP7 as the first operational implementation of the underwater robotics use case and is intended to assess both the technical performance of the AI Systems Under Test (AI SUT) and their trustworthiness in a realistic maritime infrastructure context. More specifically, the pilot addresses the use of underwater drones for inspection, maintenance support, bathymetric surveying, and anomaly detection in port areas, where limited visibility, dynamic environmental conditions, and the safety relevance of underwater operations make trustworthy AI particularly important.

The scope of the initial pilot covers two distinct but complementary AI applications developed by UNIFI. The first is Automatic Target Recognition (ATR) , deployed on the FeelHippo AUV [1], which uses Forward-Looking Sonar (FLS) data to detect, classify, and geo-localize underwater objects, anomalies, or potential threats relevant to port operations. The second is Autonomous Coverage (AC) , integrated into the reconfigurable RUVIFIST AUV [2], which autonomously computes and updates survey paths in order to achieve efficient coverage of predefined underwater areas while generating bathymetric map. Together, these two systems address two key operational needs of port infrastructure maintenance: reliable underwater perception and reliable autonomous surveying.

From an application perspective, the initial pilot is designed to demonstrate the feasibility and usefulness of AI-driven underwater robotic systems in scenarios such as the detection of submerged objects and bathymetric analysis. In this sense, the pilot is not limited to a purely technical proof of concept, but aims to evaluate how these systems can support real port activities in a safer, more efficient, and less intrusive way than traditional diver-based or vessel-based approaches. The adoption of AUVs is expected to reduce human exposure to hazardous underwater conditions, minimise operational disruption to port traffic, and improve the frequency and quality of inspections and surveys.

A central objective of the initial pilot is therefore to assess the extent to which the ATR and AC systems can operate effectively in real port conditions while satisfying the trustworthiness expectations of the FAITH project. This includes not only technical objectives such as accuracy, robustness, reliability, coverage effectiveness, and real-time usability, but also broader objectives related to safety, explainability, and operational trust. In line with the FAITH approach, the pilot seeks to bridge the gap between AI developers and end users by combining engineering validation

with stakeholder engagement, trustworthiness assessment, and iterative refinement of both the systems and their operational use.

The initial pilot also has an important methodological role within the overall FAITH project. It serves as the initial experimental basis for applying the FAITH AI Trustworthiness Assessment Framework to underwater robotics, generating the first in-isolation evidence on AI assets, human teams, threats, vulnerabilities, and preliminary countermeasures. The lessons learned from Ravenna are intended to inform subsequent improvements of the SUT and to support the preparation of the replication pilot in the Port of Valencia. In this respect, the initial pilot is both a validation activity in its own right and a foundation for the second implementation cycle of the underwater robotics LSP.

1.2. Structure of the document

This deliverable is structured to provide a coherent overview of the initial pilot phase of the underwater robotics LSP, covering both the technical progress of the AI SUT and the initial application of the FAITH trustworthiness methodology and tools.

Following this introductory section, Section 2 reports on the progress of the AI system deployment and on the execution of the pilot activities. It describes the operational context of the initial pilot, the deployment environment in the port domain, the involvement and training of users and stakeholders, the status of the replication pilot preparation, any deviations from the original deployment plan, the data collection activities carried out during the pilot, and the main lessons learned from the integration of the AI systems into the LSP workflows.

Section 3 focuses on the application of FAITH tools within the pilot. It first describes how the FAITH AI Trustworthiness Assessment Framework version 1 (FAITH AI_TAF v1) has been applied to the AI systems under assessment ATR through asset-level in-isolation assessment. It then reports on the dry-run activities, the planning of validation actions, the training delivered for the different FAITH tools, and the initial trustworthiness-related outputs obtained through tools such as trustSENSE, TrustGuard, and the FAITH AI Model Hub.

Section 4 presents the evaluation of the AI systems. This includes quantitative technical metrics collected on pilot data, as well as an initial trustworthiness assessment of the systems against the most relevant trustworthiness dimensions for the underwater robotics domain. The section also includes a preliminary qualitative evaluation of the impact of the AI systems within the specific pilot scenario.

Section 5 addresses the regulatory and ethical assessment of the pilot. It discusses the alignment of the deployed systems with the EU AI Act and other relevant frameworks, as well as local and

port-specific compliance aspects and the technical compliance of the data and AI models used in the pilot.

Finally, Section 6 summarises the main achievements of the reporting period and outlines the roadmap for the final pilot phase, including the next steps for the replication pilot and the main risks identified for the continuation of the pilot activities.

1.3. Evolution of the AI System Under Test (SUT) from initial setup

ATR

Since the initial system description reported in D7.1, the most significant evolution within the underwater robotics SUT concerns the ATR application, which has progressed from a prototype focused mainly on reliable underwater detection and risk-aware target classification to a more mature operator-centred trustworthiness-oriented decision-support system. In the initial setup, ATR was presented as an AI module integrated into the FeelHippo AUV, capable of performing instance segmentation and object detection from FLS data.

During the initial pilot cycle, and especially through the interaction with FAITH methods, end users and field constraints, this initial ATR concept has been refined in a way that makes the trustworthiness requirements much more explicit in the actual software architecture. In particular, the current application has evolved toward an operational bridge, where navigation telemetry, altitude, detections, clustered FLS detections, FLS streams, and mission-control information are exposed through a web-based interface. The updated software no longer acts only as an AI inference wrapper, but as an integrated operational layer connecting perception, mission supervision, map-based visualisation, operator commands and explainability services in a single application.

A major step forward concerns the explicit management of scenario-based ATR execution. The current application introduces dedicated support for ATR scenario selection, class-specific risk configuration, applied scenario labels, and target legends, together with dedicated handling of FLS detections and request-driven activation of sonar-based target detection. This reflects an important maturation of the system, because the AI is no longer treated as a generic detector, but as a configurable component adapted to specific mission contexts and specific categories of underwater targets. This is particularly coherent with the actual pilot evolution, where the implemented ATR validation was narrowed to the sunken-object scenario and then operationally refined through the distinction between long-range and short-range sonar inspection.

Another major evolution, strongly aligned with the FAITH trustworthiness objectives, is the strengthening of the risk-aware interaction logic. The ATR concept already foresaw that operators could define risk levels for different target classes and that the system would halt when an

unacceptable-risk target was detected, returning control to the operator. In the evolved application, this logic is implemented much more explicitly in software through dedicated unacceptable-alarm state variables, target tracking, alarm acknowledgement procedures, and mechanisms for pausing missions and requiring operator intervention. This makes the safety-oriented behaviour of the system more robust and operationally transparent than in the initial setup. In practice, FAITH has helped move the system from a conceptual statement of risk-aware behaviour to a concrete implementation of human-in-the-loop safety interlocks and operator takeover logic.

The evolution of the Graphical User Interface (GUI) is equally important. The original version already envisaged a graphical interface showing sensor outputs, overlays of detections, geo-localisation of targets, and the possibility for the operator to take control of the vehicle whenever needed. The evolved application preserves these principles but extends them with richer real-time mission status management, target snapshots, altitude telemetry, explicit mission-state emission, and web-based synchronisation of interface events. This means that the GUI is no longer only a visualisation tool, but a structured interface for situational awareness, command and control, and trust-aware supervision.

A further key improvement concerns explainability, which was one of the trustworthiness dimensions most clearly raised both in FAITH and in the stakeholder-oriented activities. ATR decisions are supported with GradCAM-style explanations and reference examples from the training dataset. This feature produces not only the Grad-CAM overlay, but also structured PDF outputs containing the model name, target class, robot position, depth and yaw, original and detected images, Grad-CAM visualisations, and examples from the training dataset.

This evolution is also visible in the practical design of the current target-detection scenario. The synthetic underwater targets prepared for the pilot campaigns, including representative objects such as a tire, a steel beam element, a human-shaped surrogate and a square concrete frame, were not only useful for validating detection performance in controlled conditions, but also contributed to making the application more operationally grounded. They enabled the ATR pipeline, the GUI logic and the explainability/reporting functions to be exercised on realistic target categories before extending the system to broader replication scenarios.

The current SUT is therefore more mature: it is more operationally integrated, because it unifies sensing, mission data and interface logic in a single application; more risk-aware, because unacceptable detections trigger explicit control and alarm procedures; more explainable, because decisions can now generate both visual and documentary outputs; and more aligned with stakeholder and FAITH feedback, because its present architecture directly reflects the trustworthiness dimensions prioritised during the pilot. The FAITH framework has been

particularly important in this evolution, because it helped transform general design intentions, such as transparency, safety, explainability and operator trust, into concrete software functionalities and control mechanisms embedded in the evolved ATR application.

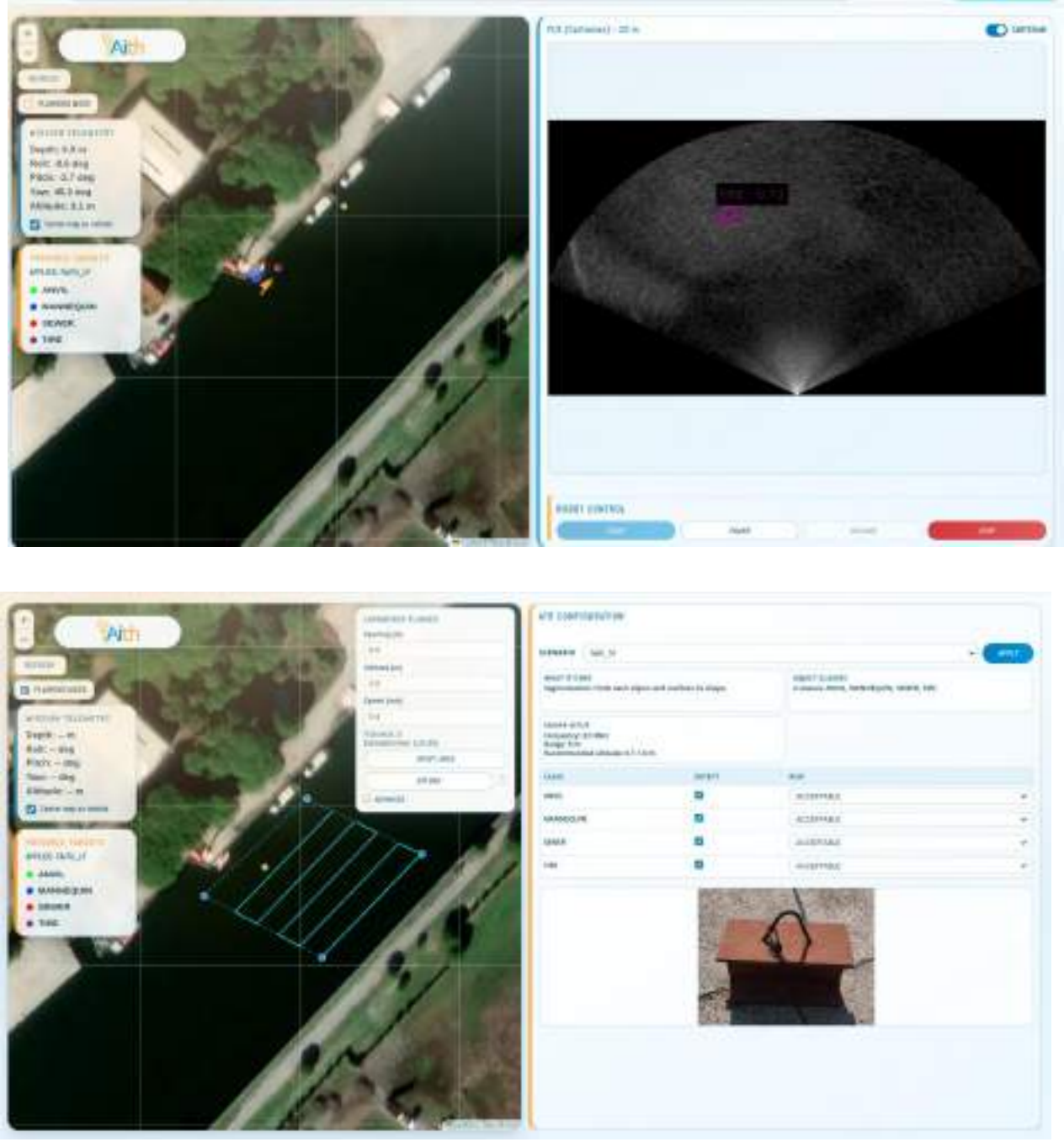


Figure 1: Example of the evolved ATR AI SUT over an experimental campaign

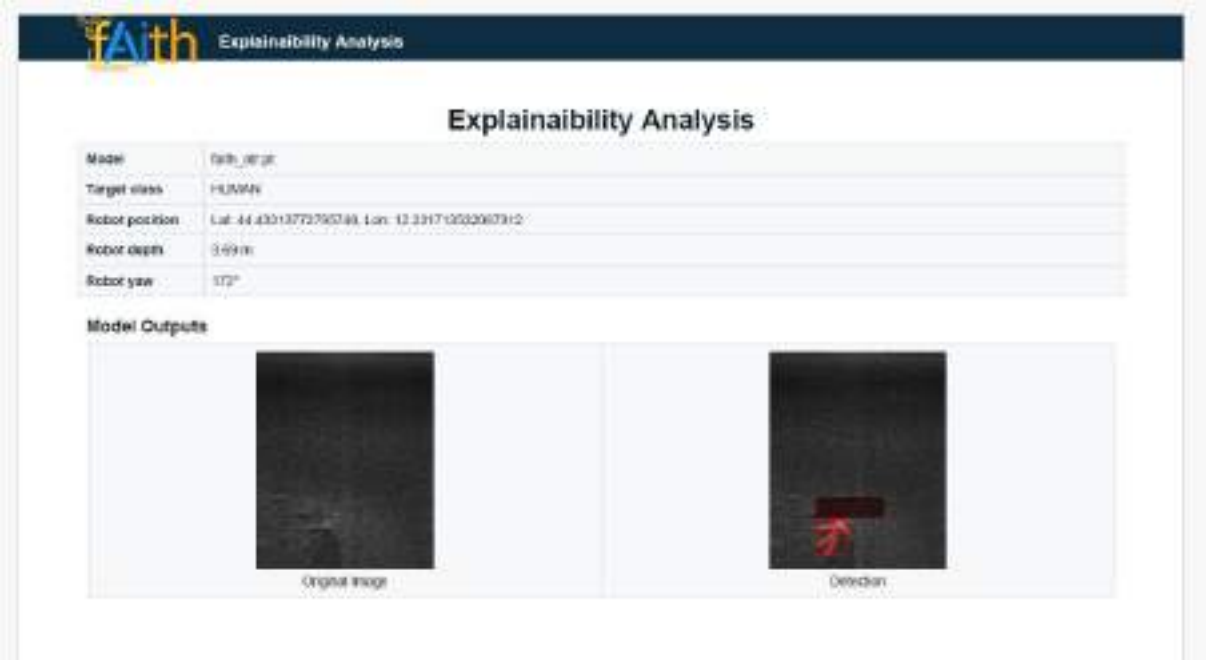


Figure 2: Example of the Explainability report per the evolved ATR AI SUT (Part 1)



Figure 3: Example of the Explainability report per the evolved ATR AI SUT (Part 2)

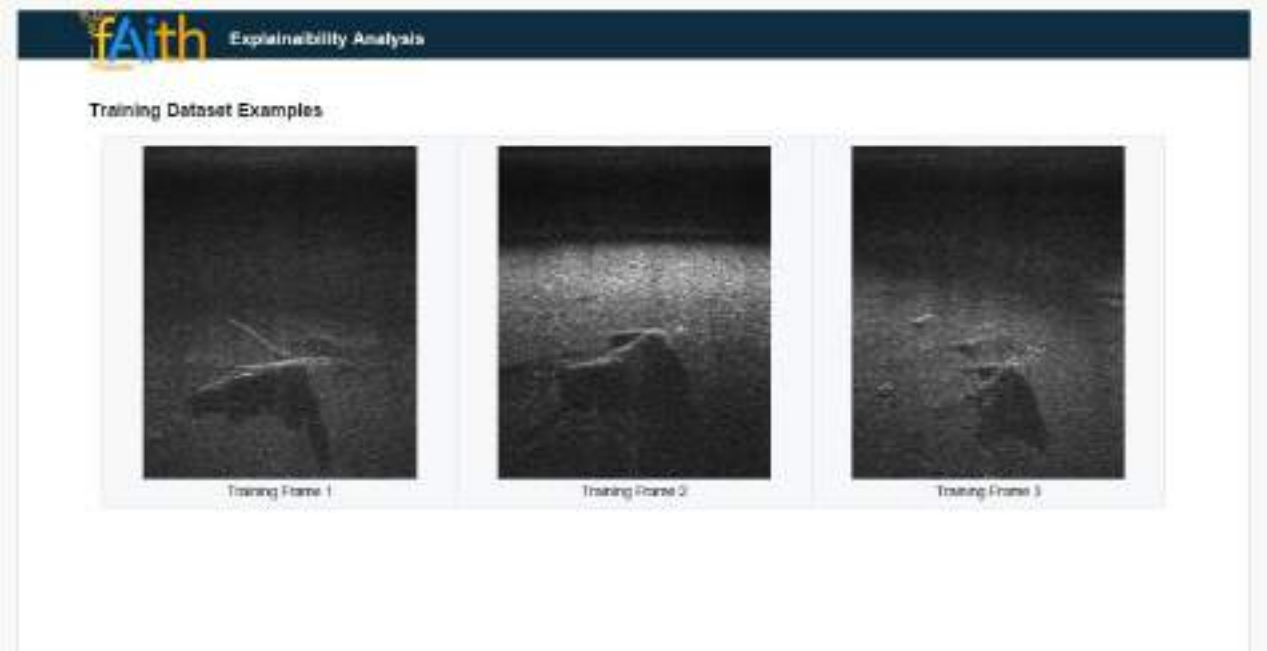


Figure 4: Example of the Explainability report per the evolved ATR AI SUT (Part 3)

AC

For the AC system, the evolution from the initial setup described in D7.1 to the current implementation is mainly visible in the transition from a planning algorithm embedded in the RUVIFIST AUV to a much more explicit human-supervised operational system, where mission definition, risk-aware replanning, and real-time situational awareness are directly exposed to the operator through a dedicated web interface. Previously, AC was presented as an AI solution integrated into the reconfigurable RUVIFIST AUV, equipped with a 3D multibeam echosounder and able to autonomously compute the optimal path for surveying a designated area once the operator had provided four or more boundary points. The system was already conceived as capable of dynamically updating the survey route according to the portion of area already covered, while continuously returning the developing bathymetry to port operators.

A first important evolution is the integration of risk-aware mission planning into the user interface. While the original version of AC already emphasised that the system should avoid obstacles, manage interdicted or partially interdicted areas, and dynamically recalculate the path, the current implementation exposes these safety-relevant concepts much more explicitly. The operator can now draw dedicated risk areas on the map, assign a risk level through a slider-based mechanism, and send those areas to the planner as structured mission constraints. This is a

substantial maturity step with respect to the initial setup, because safety is no longer only handled as an internal property of the planner, but is represented through visible, editable and auditable mission constraints.

Another relevant improvement is the increased real-time transparency of the planning process. In the current application, the generated path is transmitted back to the interface and displayed directly on the map, together with the coverage area, detected obstacles, risk polygons and a continuously updated coverage percentage. This makes the internal operation of the AC algorithm much more observable than in the original design description. Instead of simply trusting that the vehicle is surveying correctly, the operator can now verify which path has been generated, how much of the requested area has been covered, where the obstacles lie, and how the planner is progressing with respect to the intended mission. In FAITH terms, this represents a clear strengthening of accountability and transparency, because the planner's behaviour becomes visible and reviewable during execution rather than only after the mission is complete.

The evolution is also visible in the introduction of planner parameterisation at interface level. The current GUI provides two complementary parameters, α and 1−α, which are sent to the backend as explicit coverage-planning parameters. This makes the AC system more flexible and, more importantly, more aligned with the FAITH principle that AI behaviour should remain understandable and adjustable by human supervisors wherever possible. Even when the exact internal optimisation meaning of these parameters is not exposed to the user in mathematical



Figure 5: Example of the evolved AC AI SUT over an experimental campaign

terms, their presence reflects a move from a fixed autonomous behaviour to a more configurable and user-aware one.

In this sense, FAITH has contributed to the evolution of AC not by changing its core mission objective, which remains the autonomous generation of efficient bathymetric coverage, but by helping transform it into a more trustworthy system. The current implementation addresses, in a much more concrete way than the initial version, several trustworthiness dimensions that had been highlighted during the pilot, especially safety, transparency, and human oversight. Safety is strengthened through explicit obstacle and risk, area handling; transparency is improved through map-based visualisation of the area, path and progression; accountability is enhanced by making the planning constraints and mission outputs directly inspectable; and human oversight is reinforced through an interface that allows the operator to define, monitor and influence mission behaviour at runtime.

The AC SUT has evolved from an initially well, performing autonomous bathymetric planning prototype into a more mature trust-aware operational tool, better suited for deployment in the FAITH pilot context. Its current implementation shows a clear shift toward explainable mission setup, visible risk management, real-time supervision and tighter integration between autonomy and human control. This evolution reflects both the practical lessons learned during the initial pilot cycle and the influence of the FAITH methodology in translating trustworthiness requirements into concrete software and interface design choices.

2. Progress on the AI system deployment(s) and execution of the Pilots

2.1. Status of the 1st Pilot execution

2.1.1 Deployment Context & Domain Specifics

The port of Ravenna is a canal port that develops on the Candiano canal that connects the city to the northern Adriatic Sea which today has about 14.5 km of operational docks, 603,000 square meters of warehouses, 1,350,000 square meters of yards and 1,256,000 cubic meters of tanks/silos as well as areas within the port perimeter for 2,080 hectares.



Figure 6: Map of the Port of Ravenna

The Port of Ravenna is active in the field of technological innovation, particularly with the participation in many projects founded by EU Commission. Since a couple of years there have been some activities devoted to testing AI in various applications: workers safety, process optimization, and use of maritime and underwater drones. AI is being tested for optimizing vessel scheduling, autonomous navigation, mitigation of risks of collision man-machine, and automating various processes within the port. Additionally, the port is exploring the use of AI for enhancing cybersecurity, and for predictive maintenance of assets with a special focus on the maintenance of the seabed as one of the bigger problems is to maintain the necessary level of draught along the canal. These initiatives are part of the Port of Ravenna broader strategy to become a smart port.



Figure 7: Port of Ravenna area

The initial pilot was developed through a sequence of experimental campaigns carried out in two different locations, selected for complementary reasons related to environmental realism, logistics, safety and progressive system maturation.

- *Location 1*: the first location was the southern breakwater ("Diga Sud"), situated at the entrance of the port and characterised by a long breakwater configuration, approximately 8 m water depth, low visibility, and limited interference with ordinary port operations and navigation, although authorisation from the Harbour Master was required. Electricity and data connection were available on site, which made the location suitable for an initial campaign aimed at understanding the real behaviour of the sensing systems in a representative maritime environment while keeping the operational impact low.



Figure 8: "Southern breakwater" test location

Table 1: Location 1 features

Southern breakwater
Depth (m) = [8]
Low visibility
No impact in Port operations , Low impact in navigation
Safety

Harbor Master Authorization needed
Electricity and data connection available

- *Location 2*: the second location was the "Firefighters Port Base", the section of the canal in front of the Firefighters Port Base, a section of the port canal with low level of traffic and the possibility to use the facilities of the Firefighters for launch and hauling of the vehicles and for all the practical needs. This site, which became the main operational base for the later campaigns, was particularly useful because it enabled close cooperation with the nautical unit of the Fire Brigade, simplified the repeated deployment of targets and vehicles, and supported a more direct interaction with the end users who would later operate and assess the systems in practice.

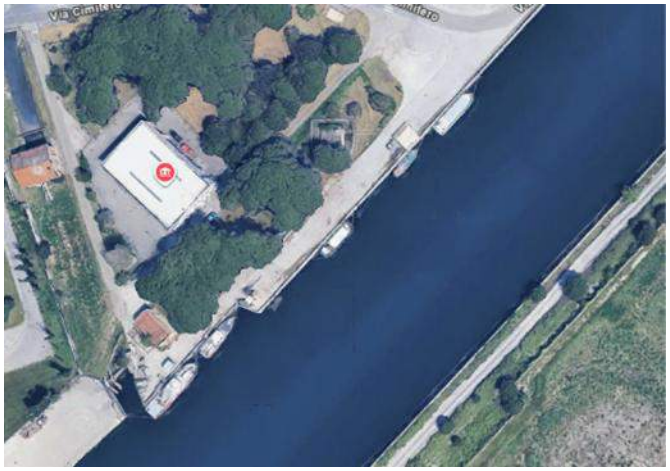


Figure 9: "Firefighters Port Base" test location

Table 2: Location 2 features

City dock
Depth (m) = [7]
Very low visibility
Low impact in Port operations
Safety

No Authorization needed
Firefighters' premises available for electricity, water, storage of devices, ...

Experimental campaigns

The first experimental campaign, carried out in Location 1 from 5 to 7 May 2025, was mainly devoted to contextualising the sensing payloads of the vehicles with respect to the actual objectives of the pilot. This phase was essential to understand the concrete perceptual limitations of the ATR system in real underwater conditions, especially with respect to visibility, target observability and scenario feasibility. During this campaign, representative targets were also tied to floating buoys in order to make their position known and facilitate controlled observation trials. The objective was to verify whether such targets could be effectively observed with the sensing payloads required by the pilot, in particular through the FLS mounted onboard the FeelHippo AUV. In parallel, the campaign also explored the possibility of observing quay damages and submerged infrastructures, in order to assess the practical feasibility of infrastructure-inspection scenarios under the actual environmental conditions of Ravenna. It also allowed to clarify which scenarios were realistically executable in Ravenna and to start a first campaign of data collection.

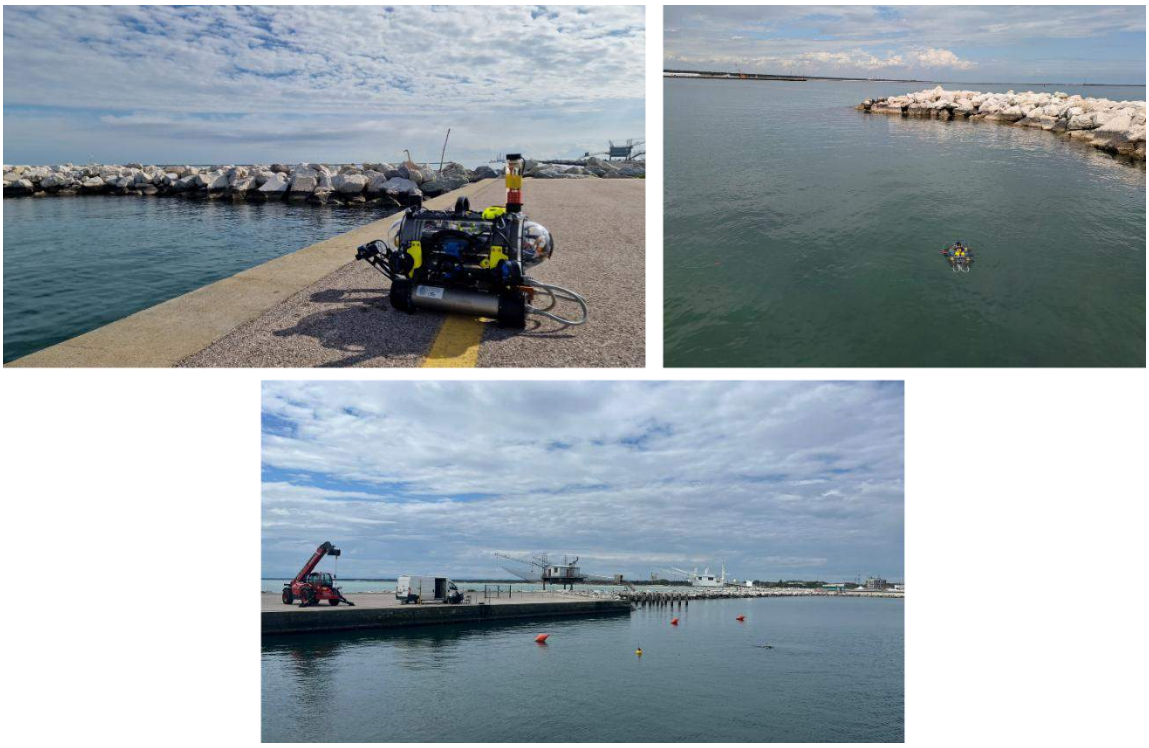


Figure 10: Visual evidences of the first experimental campaign

In practical terms, this campaign acted as an exploratory stage, bridging the gap between the intended pilot scenarios and the actual environmental constraints imposed by the port domain.

The second experimental campaign, carried out in Location 2 from 24 to 26 June 2025, focused on structured data collection for ATR in a more heterogeneous environment. During this phase, specific representative targets were defined and positioned on the seabed in order to support the training of the models. These targets were tied to floating buoys so that their position could be more easily known and controlled, thus facilitating repeatable data collection and annotation. This campaign therefore had a clear dataset-building function, allowing the consortium to generate a more suitable training corpus for the target, recognition task. On the AC side, the same phase also included first test missions aimed at checking the correct operation of the AC chain before the full deployment of the bathymetric configuration.

The third experimental campaign, carried out again in Location 2 from 23 to 27 March 2026, marked the first real transition from data preparation to testing with trained ATR models and to the use of the new version of the ATR AI SUT. This phase included the deployment of ATR for use



Figure 11: Illustrations of the second experimental campaign



Figure 12: Visual reports of the third experimental campaign

by the nautical detachment of the Ravenna Fire Brigade, thus moving the pilot from internal technical validation toward first operational use by an actual end-user organisation. In parallel, the AC activities progressed with the first consolidation tests related to the newly acquired bathymetric sensor. This campaign was therefore relevant both from the technical perspective, because it enabled the first validation of the trained models and updated ATR application, and from the operational perspective, because it introduced real users into the loop.

The fourth experimental campaign, also carried out in Location 2 from 20 to 22 April 2026, represented a second and more advanced testing stage for both applications. For ATR, the trained models and the evolved ATR application were tested again with the involvement of both the Ravenna Fire Brigade and the APRA. For AC, this campaign marked the deployment of the evolved application for direct use by the same two end-user organisations. In addition to route planning and mission supervision, this phase also produced bathymetric outputs as a direct result of the AC algorithm, showing a higher level of maturity of the AC chain in realistic port conditions. This final campaign therefore demonstrated a clear progression from exploratory sensing and data collection toward real pilot use of the evolved ATR and AC applications by operational stakeholders.

From a domain-specific point of view, the deployment context imposed several relevant constraints. Underwater visibility ranged from low to very low depending on the selected area,

which strongly influenced the sensing strategy and reinforced the importance of sonar-based perception for ATR and multibeam-based mapping for AC. Logistics and authorisation requirements also varied between sites: some areas allowed low-impact testing with limited interference to navigation, whereas others required Harbour Master approval. At the same time, the availability of shore-side infrastructure such as electricity, connectivity, storage space and operational support was an important factor in enabling repeated deployments and interaction with end users. These aspects were not secondary conditions, but structural elements shaping the entire execution of the initial pilot.

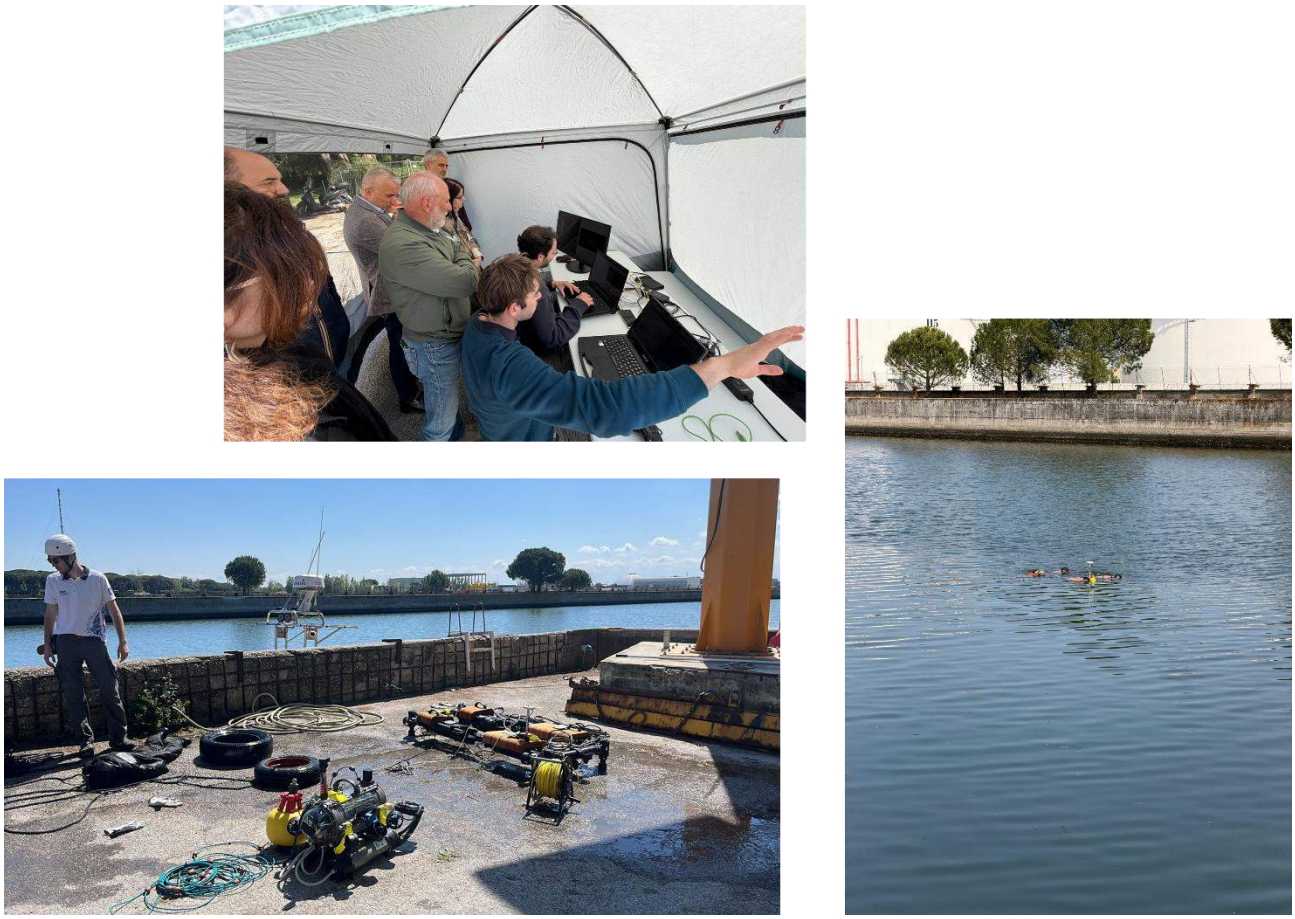


Figure 13: Visual reports from the fourth experimental campaigns

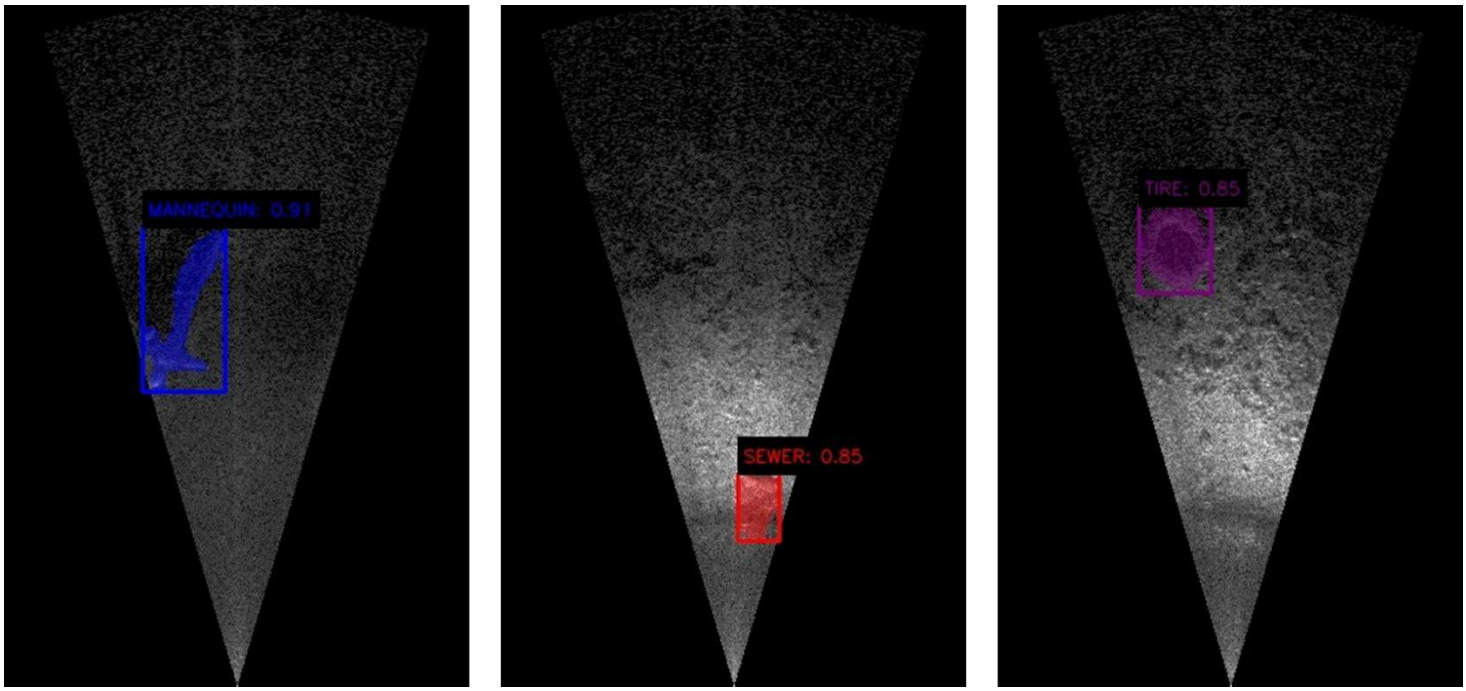


Figure 15: Visual evidences of ATR SUT deployment on high-frequency FLS frames over the third and fourth experimental campaign

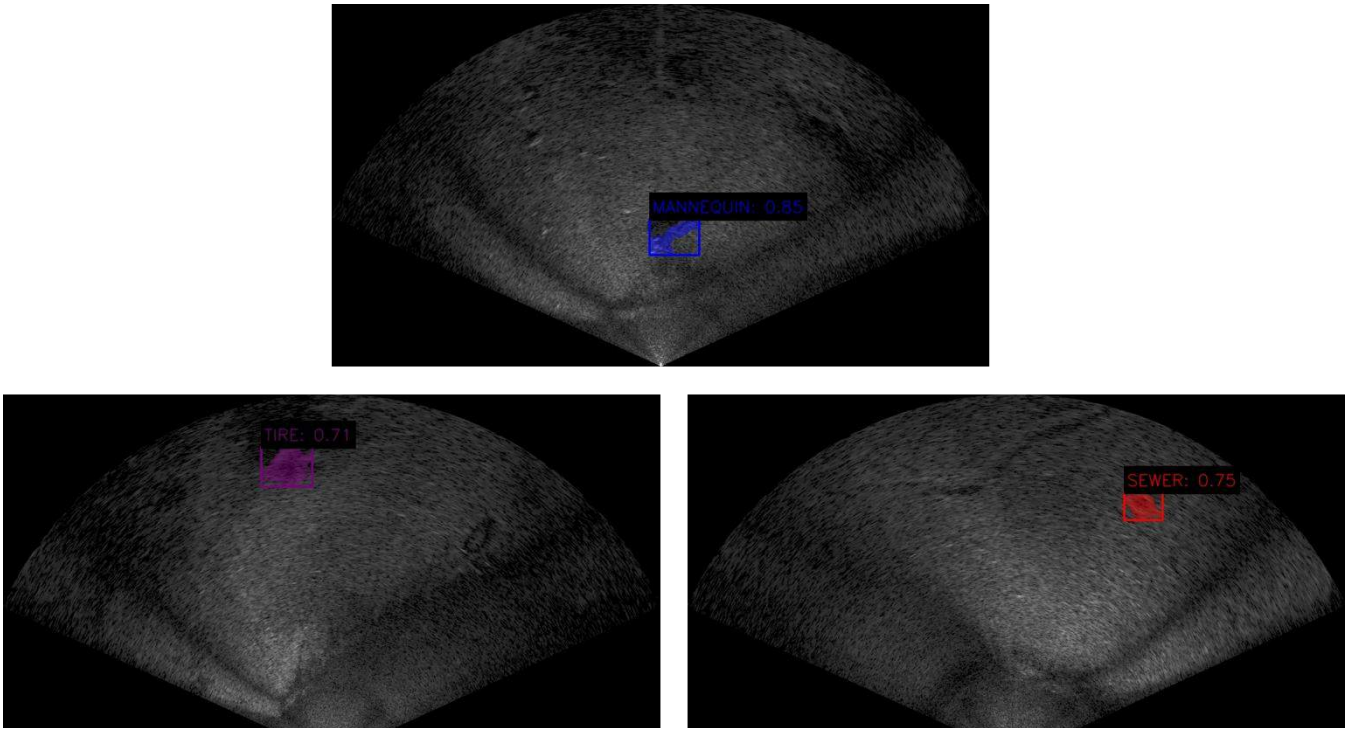


Figure 14: Visual evidences of ATR SUT deployment on low-frequency FLS frames over the third and fourth experimental campaign

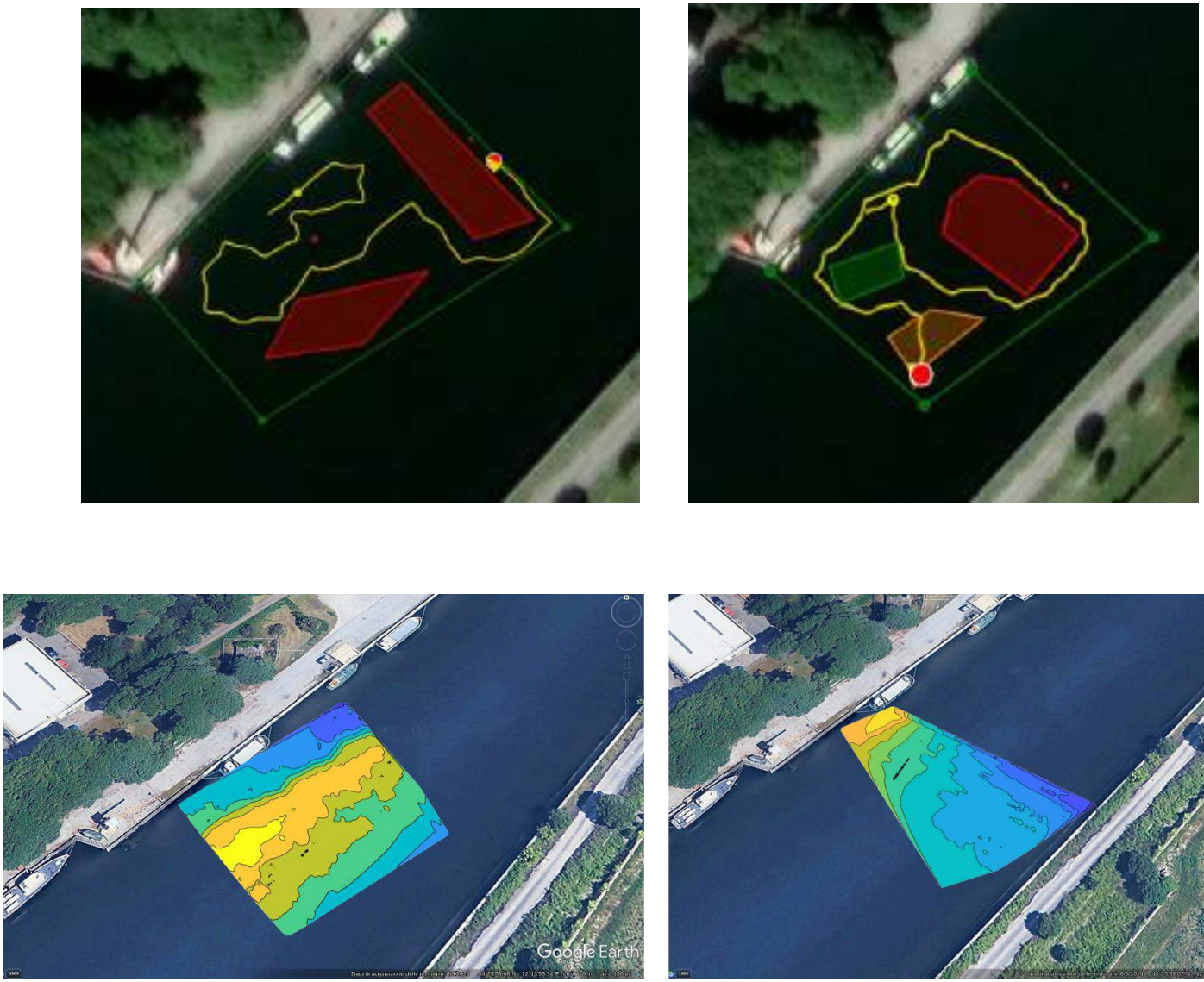


Figure 16: Visual evidences of AC SUT deployment on validation areas over the third and fourth experimental campaign

2.1.2 Training

End-user training constituted a key enabling activity for the initial pilot, as the successful deployment of the AI SUT depended not only on the technical maturity of the ATR and AC solutions, but also on the ability of the relevant operational actors to understand their purpose,

limitations, expected benefits, and modes of interaction. In line with the user engagement strategy already defined in D7.1, the training activities were conceived as a progressive process combining early workshop-based familiarisation, interactive discussion with stakeholders, and subsequent hands-on field training with the vehicles and their operator interfaces. This approach was adopted to ensure that the end users were not exposed to the systems as black-box technologies, but were gradually introduced to the pilot rationale, the specific AI functionalities, and the human oversight mechanisms foreseen for their practical use.

First workshop

A first structured training and awareness activity was carried out through the Ravenna workshop held on 4 March 2025, which represented the initial common entry point for end users and stakeholders into the underwater robotics pilot. The event, attended by 23 to 27 participants both in person and remotely, was explicitly designed as a bridge between AI, underwater drones and the ports of the future, and had a dual objective: first, to provide a technical and strategic overview of the FAITH project and the underwater robotics pilot; second, to create an interactive environment in which port operators, institutional actors, academic experts and other stakeholders could discuss requirements, expectations and trustworthiness concerns related to the proposed AI systems. The agenda included an introduction to the project and the pilot, a technical deep dive on the underwater drones and AI systems involved in the study, a presentation of the port authority's perspective on the requirements for trustworthy AI, and a dedicated session on the solutions proposed by the developers to improve AI trustworthiness. This first workshop had an important training function for end users because it did not limit itself to a general project presentation. Rather, it introduced the participants to the concrete operational logic of the SUT. The technical session on the pilot explained the available AUVs, their sensing payloads, and the two AI applications under evaluation, namely ATR and AC. In particular, the participants were introduced to the FeelHippo AUV and its FLS for ATR, as well as to the RUVIFIST vehicle and the multibeam-based bathymetric surveying capability supporting AC. The presentation material also explained the role of the operator interfaces, the communication architecture, and the human-on-the-loop paradigm adopted in the pilot, whereby the operator remains able to monitor the system and intervene whenever necessary. This was especially relevant for end users from the port domain, as it helped clarify from the outset that the systems were intended to support operational decision-making rather than replace human control.

Another important element of the workshop-based training was its interactive component. The agenda explicitly included a session dedicated to stakeholder discussion around three core themes: explainability and comprehension of ATR decisions, reliability and robustness of ATR, and safety of AC when the robot autonomously decides where to go. A final brainstorming session was used to identify which aspects still needed to be strengthened in order to make the proposed

AI systems truly trustworthy from the perspective of end users. This structure was particularly valuable because it transformed the workshop into a two-way training process. On the one hand, the technical team explained the systems, their trustworthiness objectives, and the adopted mitigation strategies; on the other hand, the operational stakeholders could express what they expected from the systems in terms of interpretability, reliability, intervention capability and safety guarantees. In this sense, the workshop served not only to inform end users, but also to calibrate the training and validation process around their real informational and operational needs.

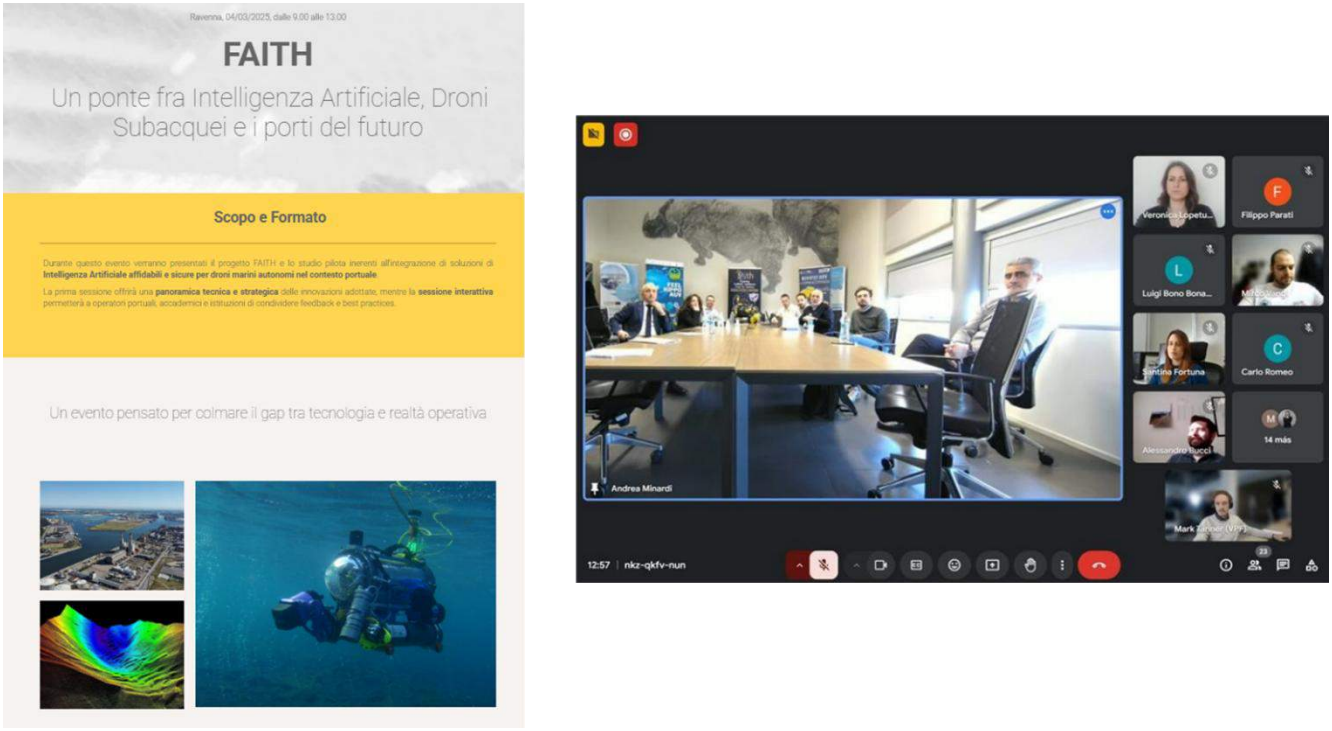


Figure 17: Visual reports from the first workshop

Practical sessions

Following this initial workshop, the end-user training continued through on-field, in-person practical sessions carried out in connection with the deployment activities. These sessions were essential to move from conceptual understanding to operational familiarity. In practical terms, the field training focused on enabling the relevant end users to observe and understand the complete deployment workflow of the underwater systems, including vehicle preparation, sensor setup, communication checks, mission planning, mission supervision, and interpretation of

system outputs. These activities involved direct interaction between the technical team and the operational actors at the deployment site, thereby allowing users to become familiar with the actual hardware platforms, the shore-side setup, and the sequence of actions required to conduct an inspection or surveying mission.

For the ATR component, the field training exposed end users to the operational use of the FeelHippo AUV, the role of the FLS-based perception pipeline, and the way in which detections are presented to the operator. The training addressed the communication architecture based on the passive buoy and the real-time visualization of outputs, the interpretation of detections and geo, localisation on the operator interface, and the logic through which suspicious or risky targets are highlighted and can trigger operator attention or intervention. The training also familiarised users with the explainability-related outputs and with the supervision logic under which the operator can monitor the mission and take control when needed. This was particularly important for end users such as the APRA and the Fire Brigade, because the practical value of ATR depends heavily on their ability to correctly interpret detections and integrate them into their operational response process.

For the AC component, the field training addressed the use of the GUI for selecting the survey area, defining boundary points, setting the expected coverage, and monitoring the progression of the bathymetric mission. End users were shown how the AC functionality computes the route, how the system updates the surveyed area in real time, how bathymetric outputs are visualised, and how operators can maintain supervision throughout the mission. In addition, the training reinforced the human-on-the-loop logic already presented during the workshop, by showing in practice how the operator can stop the mission or switch from autonomous to teleoperated control if required. In this way, the field sessions helped translate the abstract concept of trustworthy autonomy into concrete operational behaviour observable by end users.

Beyond the transfer of procedural knowledge, the field training also had an important trust-building function. Seeing the vehicles, communication setup, interfaces and AI outputs in real deployment conditions allowed the end users to move from a theoretical understanding of the systems to a more grounded perception of their capabilities and limitations. This was especially relevant in a port context, where trust in autonomous systems depends strongly on the visibility of safeguards, on the possibility of human intervention, and on the clarity with which the system communicates its outputs. The combination of technical explanation and direct exposure to the deployed systems therefore contributed to building a shared operational understanding between developers and users.



Figure 18: Visual example of on-field training activity

Second workshop

An additional workshop/training session was then carried out on 20 April 2026, with the specific objective of updating the APRA and the Fire Brigade on the progress of the AI SUT and explaining the updated system behaviour before asking them to use the solutions in practice. This additional session, attended by 11 participants, was particularly important because the pilot had evolved between the first awareness workshop and the actual operational testing. The session therefore served as an update and re-alignment activity, ensuring that the end users were informed about the developments introduced in the AI SUT, the refined operating logic, and the updated interaction procedures before engaging with the systems in the field. In this sense, this session should be understood as a second, layer training action, aimed not at first familiarisation, but at operational readiness and informed use of the updated pilot systems.



Figure 19: Visual reports of the second workshop/in-depth training session

Summary

The end-user training process was therefore not a single isolated event, but a staged activity composed of an initial stakeholder workshop, subsequent field-based hands-on training, and an additional update session immediately preceding use of the evolved systems. This progression ensured that end users were first introduced to the concepts and objectives of the pilot, then exposed to the systems in real deployment conditions, and finally updated on the latest developments before practical use. Such a training strategy is fully coherent with the human-centred and trustworthiness-oriented approach of the FAITH project, in which meaningful adoption of AI systems in critical environments requires not only technical performance, but also informed, supervised and confident use by the relevant operational actors.

2.1.3 Users, Stakeholder identification and involvement

The identification and involvement of users and stakeholders was treated as a core component of the initial pilot, because the underwater robotics use case does not depend only on the technical maturity of the ATR and AC systems, but also on their acceptance, usability and operational relevance for the port community. The stakeholder strategy adopted for the pilot was therefore based on a structured combination of stakeholder mapping, outreach and engagement, feedback collection, training and support, regular updates, and joint evaluation, with the explicit objective of ensuring that the deployed AI-enabled underwater robotic systems responded to real operational needs and trustworthiness expectations. This approach was already defined in the pilot-design phase and was then carried forward into the actual execution of the Ravenna pilot.

A first step consisted of stakeholder mapping, aimed at identifying the actors that could influence, use, supervise, regulate, or be affected by the pilot activities. In the port domain, this mapping included institutional stakeholders, operational end users, technical partners, and external expert bodies. Among the main actors considered were the Port Authority departments involved in industrial safety, operational safety and security, the Harbour Master and maritime regulatory authorities, the Ravenna Fire Brigade, and the technical and scientific partners contributing to the development and trustworthiness assessment of the systems, including UNIFI, Interuniversity Center of Integrated Systems for the Marine Environment (ISME). The purpose of this mapping exercise was not only to list actors, but to clarify their role, their expectations, and the specific way in which they could contribute to the pilot.

Within the actual initial pilot, the most directly involved end-user organisations were the APRA and the nautical detachment of the Ravenna Fire Brigade, which progressively moved from stakeholder consultation to practical interaction with the SUT. The experimental campaigns were deliberately organised so that the pilot evolved from initial technical contextualisation to a stage in which these actors were directly involved in testing the updated ATR and AC applications in realistic field conditions. This progression was important because it ensured that end-user involvement was grounded in real system behaviour rather than limited to abstract discussion.

To support active participation, the pilot adopted a strategy of outreach and engagement based on both collective and direct interaction formats. A first central milestone was the Ravenna workshop, which created a common entry point for stakeholders and end users and enabled direct discussion of the pilot objectives, the underwater robotics solutions under development, and the trustworthiness requirements associated with their operational adoption. Alongside this collective engagement format, more direct one-to-one interactions and practical exchanges were used during pilot preparation and execution to align expectations, clarify use cases, and prepare the operational actors for later field use. In this way, engagement was not treated as a one-off dissemination action, but as a continuing process of alignment between developers, pilot coordinators and users.

A further important dimension was the use of feedback mechanisms. The pilot strategy explicitly foresaw the use of interviews, focus groups and other structured discussion formats to gather input on the performance, usability and trustworthiness of the underwater robotic systems. In practice, the workshop and subsequent interactions were used to collect stakeholder views on topics such as the interpretability of ATR outputs, the robustness and reliability of the systems, the safety of autonomous operation, and the adequacy of human oversight mechanisms. This was especially relevant in a critical port environment, where system acceptance depends heavily on whether users feel able to understand outputs, supervise behaviour, and intervene when

necessary. The feedback collected through these mechanisms was therefore not peripheral, but part of the evidence base used to refine the systems and their operational deployment logic.

The pilot also placed strong emphasis on training and support, because meaningful user involvement requires more than consultation: it requires operational familiarity with the systems. The training activities, described in the preceding subsection, were therefore directly connected to the stakeholder-involvement strategy. They ensured that the main end users were gradually introduced to the rationale of the pilot, to the purpose and limitations of the ATR and AC systems, and to the human-on-the-loop mechanisms through which the systems are meant to be supervised in practice. This training dimension was essential to transform stakeholders from passive observers into informed users and evaluators of the deployed solutions.

Another important element of the involvement strategy was the commitment to regular updates. Stakeholders and end users were kept informed about the progress of the pilot, the evolution of the SUT, and the main adjustments introduced during the execution of the campaigns. This was particularly important because the pilot was not static: both ATR and AC evolved over time, and users needed to remain aligned with the updated system behaviour before being asked to interact with the solutions in the field. In this sense, regular updates functioned as a trust-building mechanism, helping to preserve transparency and continuity in the relationship between the technical team and the operational community.

Finally, the pilot adopted a logic of joint evaluation, in which stakeholders and end users were not only consulted, but also involved in assessing the performance and usefulness of the underwater robotic systems. This joint-evaluation dimension became more concrete in the later campaigns, when the evolved ATR and AC applications were used with the involvement of the APRA and the Fire Brigade. Through this process, the pilot could assess not only whether the systems were technically functional, but also whether they fit operational workflows, supported user decision-making, and met the practical expectations of the port community. This is particularly important in the FAITH context, because trustworthiness is not reduced to technical performance alone, but also includes acceptability, clarity, safety perception and usefulness in real operations.

2.1.4 Evaluation Process

The evaluation process for the initial pilot can be understood as a multi-layered assessment framework. APRA contributed the pilot, side evaluation setup by framing the systems in real port operations, involving end users, and defining the operational perspective from which usefulness and acceptability were observed. UNIFI contributed the technical evaluation methodology by defining the quantitative metrics, the scenario-specific validation logic, and the system-level criteria used to assess ATR and AC performance. The result is a coherent process in which technical evidence, operational realism, and stakeholder feedback are combined to generate a

preliminary but meaningful evaluation of the underwater robotics SUTs in Ravenna. This same evaluation logic will also support the next pilot cycle, where the replication phase will allow the consortium to extend and further validate the evidence collected so far.

The evaluation setup was organised around the progressive execution of the Ravenna campaigns and around the direct involvement of the relevant operational stakeholders, in particular the APRA and the nautical detachment of the Fire Brigade. In practical terms, the evaluation process was structured so that the systems were first contextualised and technically matured through preparatory campaigns, and only afterwards exposed to operational users in field conditions. This made it possible to evaluate not only the raw outputs of the AI systems, but also their usability, their consistency with the expected operational workflows, and their capacity to support underwater inspection and bathymetric surveying in a way that fits real port needs. The process therefore included observation of missions, supervision of system behaviour during deployment, interpretation of outputs by end users, and collection of direct feedback on clarity, usefulness, intervention capability and overall confidence in the systems. This approach is coherent with the broader pilot methodology, which explicitly relies on stakeholder engagement, training, feedback mechanisms and joint evaluation as core components of the LSP.

The evaluation process was also shaped by the intended operational role of the two systems. For ATR, the assessment focused on the ability of the FeelHippo AUV and its perception pipeline to detect underwater objects of interest, classify them correctly, and communicate the relevant information to the operator in a timely and understandable way. This includes not only the correctness of the detection itself, but also the clarity of the interface, the management of potentially risky detections, the usefulness of the geolocalisation output, and the degree to which the operator can interpret and act upon the system's findings. For AC, the evaluation focused on the ability of the RUVIFIST AUV and its coverage-planning functionality to autonomously generate and update a suitable route, achieve the requested coverage of the selected area, provide understandable bathymetric outputs, and allow the operator to supervise and, when necessary, intervene in the mission. In both cases, the evaluation was therefore intentionally designed to assess the systems not as isolated algorithms, but as deployed applications embedded in operational workflows involving human supervision.

From the technical perspective, the evaluation methodology relied on the KPI and expected, outcome framework previously defined for the underwater robotics pilot. For ATR, the technical evaluation addressed three main dimensions: detection performance, operational efficiency, and robustness/adaptability. Detection performance was assessed through metrics such as precision, recall, mean Average Precision, false positive rate and false negative rate. Operational efficiency was evaluated through detection latency and processing throughput, while robustness was

assessed by considering the system's performance in low-visibility conditions, which are especially relevant in Ravenna.

For AC, the technical evaluation methodology focused on coverage performance, operational efficiency, and mission safety/reliability. The main quantitative dimensions considered were coverage completeness, survey duration, data resolution, mission success rate and real-time processing latency. In addition, operational cost efficiency and safety improvement remained part of the overall evaluation logic, even where they could only be estimated qualitatively or preliminarily during the initial pilot.

A key characteristic of the evaluation process is that it was conceived as iterative rather than one-off. The initial campaigns were used to understand environmental and sensor limitations, refine datasets, and improve the systems before their exposure to operational users. The later campaigns then allowed the consortium to evaluate the updated AI systems in more realistic conditions and to gather new evidence that could feed further refinements. In this sense, evaluation was embedded directly into system evolution. This is particularly visible in the way the Ravenna pilot progressed from exploratory contextualisation, to structured data collection, to first testing, and finally to the use of the evolved ATR and AC applications by end users. The evaluation process therefore acted not only as a measurement exercise, but also as a mechanism for guiding technical adjustments and improving the trustworthiness of the deployed solutions.

Another important component of the evaluation process was the explicit inclusion of user and stakeholder judgement. The pilot does not assess success only in terms of whether the AI system computes the correct output, but also in terms of whether port actors and operational users can understand, trust and meaningfully use the output. This is why the evaluation process was connected to workshops, field training, live demonstrations and hands-on use by the Fire Brigade and the Port Authority. Through these activities, the consortium could evaluate dimensions such as perceived clarity of the dashboard, interpretability of detections and bathymetric outputs, adequacy of operator intervention mechanisms, and overall fit with port workflows. These user-centred aspects were essential because both AI applications operate in critical environments where trust and supervision are central to adoption.

The evaluation process was also aligned with the broader trustworthiness assessment logic of the project. Beyond classical technical metrics, the systems were evaluated against relevant trustworthiness dimensions such as safety, validity and reliability, explainability and interpretability. In practical terms, this means that the pilot evaluation considered not only whether the system performed well, but also whether it behaved in a way that supported safe and understandable decision-making. For ATR, this included attention to explainability outputs, risk-oriented detections and operator takeover mechanisms. For AC, it included attention to route

transparency, mission supervision, explicit handling of obstacles and risk areas, and the visibility of coverage progression and bathymetric outputs. This broader interpretation of evaluation is fully coherent with the FAITH objective of assessing AI systems not only as technical artefacts, but as operational tools whose trustworthiness is central to their acceptance and usefulness.

2.2. Status of the Replication Pilot preparation/initialization

As of March 2026, the preparation of the replication pilot in the Port of Valencia (Spain) is progressing according to the FAITH project's WP7 framework. This second pilot phase aims to validate and refine the AI-driven underwater robotic solutions deployed during the initial pilot in Ravenna, with a focus on enhancing trustworthiness, operational reliability, and scalability. The pilot is designed to build upon the lessons learned from the first phase, ensuring continuous improvement through iterative evaluation and adaptation.

Significant progress has been made in stakeholder engagement and user recruitment. The identification and initial recruitment of end-users have been completed, aligning with the first two steps of the 3, step methodology outlined in D7.1. The third step, which involves active user engagement, is currently in progress, ensuring that participants are fully integrated into the pilot's objectives and operational requirements.

Regarding technical and operational readiness, the infrastructure for the replication pilot is in the finalization phase. While the deployment of AUVs and their associated AI systems is underway, no testing has been conducted in Valencia to date. The focus remains on ensuring full compatibility with the port's operational environment and addressing any logistical challenges that may arise.

Regulatory and ethical compliance has been carefully assessed. No ethical or regulatory barriers have been identified that would impede the pilot's execution. The only outstanding action is the formal request for permits from the port authority, which will be addressed prior to the pilot's commencement.

Training and capacity building for end-users and technical staff are planned throughout 2026. These sessions will ensure that participants are fully prepared to operate the updated AI systems, adhere to operational protocols, and contribute to the pilot's success.

2.3. Deviations from the original deployment plan

Compared with the original deployment plan defined in D7.1, the implementation of the initial pilot required some targeted adaptations to the set of scenarios actually executed in Ravenna. In the original design, the ATR system was intended to be deployed in three distinct operational scenarios. The first one concerned the detection of sunken objects and/or subsea intruders, with

the goal of identifying and classifying submerged hazardous debris or unauthorized underwater vehicles that could represent a threat to port operations. The second scenario concerned the inspection of underwater infrastructure, where the system was expected to analyse submerged port structures such as docks, pipelines, and support elements in order to detect visible anomalies or structural damage and support maintenance prioritisation. The third scenario concerned on, demand ship hull inspection, where the system was planned to navigate alongside a vessel and identify defects or anomalies on the hull surface in real time, so as to support preventive maintenance and reduce vessel downtime. For the AC system, the original deployment plan included two scenarios. The first one, bathymetry analysis, was aimed at producing detailed and accurate seabed maps of port areas, with the additional possibility of identifying depth irregularities or formations potentially affecting navigation or port operations. The second one, change detection, was conceived as a repeated, survey scenario in which newly acquired bathymetric data would be compared against previous surveys in order to detect relevant seabed changes over time, whether caused by natural processes or by man, made interventions.

During the preparation and execution of the initial pilot, however, it became necessary to revise the practical scope of these scenarios in order to maintain a technically credible and operationally meaningful demonstration. For ATR, the pilot was ultimately limited to the first original scenario, namely the detection of sunken objects. The reason for this deviation is that, in the actual deployment conditions, no sufficiently evident and demonstrable anomalies were available in the inspected infrastructures, nor were suitable hull conditions identified that could support a robust validation of the infrastructure, inspection and ship, hull, inspection scenarios. In other words, while these scenarios remained relevant from a use, case perspective, the available real-world conditions in the initial pilot did not provide sufficiently clear target events for a reliable and convincing demonstration of those two functionalities.

To address this limitation without compromising the original purpose of the pilot, a realistic but controlled variant of the first ATR scenario was implemented. More specifically, a fictitious underwater target-detection scenario was recreated by deploying synthetic targets in the water with the support of the nautical unit of the Fire Brigade. This allowed the consortium to preserve the core objective of the original ”detection of sunken objects and/or subsea intruders” scenario, namely the automatic recognition of submerged objects of interest in a port environment, while ensuring that the objects to be detected were clearly available and operationally suitable for the validation campaign. The current ATR deployment therefore focuses on the automatic recognition of underwater targets by the vehicle through its FLS payload, rather than on naturally occurring anomalies in infrastructures or ship hulls.

In addition, the ATR deployment concept was refined with respect to what was originally envisaged in D7.1. While the original scenario already assumed the use of sonar and optical

sensing for object recognition, the actual pilot introduced a more operational inspection logic based on two complementary sonar configurations. A long-range inspection mode was added through the use of a lower, frequency FLS, allowing operators to scan a wider area and identify possible anomalies or targets from a greater distance. This was complemented by a short-range inspection mode-based on a higher, frequency and therefore higher, resolution FLS, enabling the vehicle to approach the suspected target and carry out a more detailed close-range inspection. This refinement better reflects a realistic operational workflow in low-visibility port surveillance and inspection, where a broad initial search is followed by a focused confirmation phase.

For AC, the deviation concerned the exclusion of the second original scenario, namely change detection, from the initial pilot. In D7.1, this scenario had been conceived as a repeated mapping operation in which the AUV would compare new bathymetric data with historical maps and automatically highlight relevant changes in the seabed. In practice, however, the execution of this scenario would have required clearly visible and operationally significant bathymetric differences between a reference survey and a new survey. Under the environmental and logistical conditions of the initial pilot, such variations could not be expected at a sufficient magnitude to support a meaningful validation of the change, detection functionality. For this reason, the AC deployment was restricted to the first original scenario only, namely bathymetry analysis, with a focus on AC, path generation, real-time survey execution, and production of bathymetric maps in realistic port conditions.

These adaptations should not be interpreted as a reduction of the ambition of the pilot, but rather as a necessary alignment between the original plan and the actual feasibility conditions of the first deployment campaign. The selected deviations allowed the consortium to concentrate on those ATR and AC functionalities that could be evaluated in a controlled, representative, and technically robust manner, while preserving coherence with the intended use cases described in D7.1. At the same time, it is important to note that all the original scenarios described in D7.1 remain relevant and will be kept under consideration for the replication pilot, provided that the logistical conditions, site accessibility, authorisations, and environmental characteristics of the test areas make their execution feasible. This means that the infrastructure, inspection and ship, hull, inspection scenarios for ATR, as well as the change, detection scenario for AC, may still be reintroduced in the replication phase if the Port of Valencia or other suitable testing conditions allow for a more representative and operationally meaningful validation of those use cases.

2.4. Data collection throughout the 1st Pilot execution

Data collection throughout the initial pilot execution was a core enabling activity for the ATR branch of the underwater robotics LSP, because the development of the perception pipeline required the creation of dedicated FLS datasets acquired directly in the Ravenna pilot

environment. The overall strategy followed the methodological approach defined for ATR, namely a combination of on-field acquisition, expert annotation, use of pre-trained instance, segmentation models, and extensive data augmentation, adopted specifically to address the scarcity and heterogeneity of underwater labelled data.

For ATR, the collection campaign was organised around two distinct FLS operating configurations, which later corresponded to the two CNN trained models used in the pilot. A low-frequency FLS dataset was created to support wider, area and longer, range inspection, thus capturing data suitable for first target localisation and broader search. In parallel, a high-frequency FLS dataset was collected to support closer, range and higher, resolution inspection, in line with the operational need for more detailed target confirmation once a suspicious object has been identified. The resulting dataset sizes confirm the relevance of this effort: the high-frequency dataset contains 2489 training images, 533 validation images, and 533 test images, while the low-frequency dataset contains 1428 training images, 306 validation images, and 307 test images. The collected data were organised around a well, defined target taxonomy derived from the model configuration. In particular, the ATR segmentation setup explicitly defines four classes: ANVIL, MANNEQUIN, SEWER, and TIRE, and the dataset folders are structured into dedicated train, validation and test partitions for the high-frequency branch. This confirms that the pilot did not collect generic sonar imagery, but targeted data intended to support supervised instance segmentation on a controlled set of underwater objects of interest.

A crucial part of the data pipeline was the annotation phase. In line with the ATR design guidelines, the sonar frames were transformed into supervised-learning samples through manual polygon-based instance, segmentation annotation, where each target was outlined precisely and assigned the corresponding class label. This annotation strategy is particularly important for FLS imagery, because the ATR task is not limited to image, level recognition but requires the model to learn the shape and location of the target within the sonar frame. The use of polygon masks therefore allowed the collected raw acquisitions to be turned into training, ready labelled data suitable for fine, tuning an instance, segmentation architecture.

The field, acquisition campaign also followed a set of general FLS, specific collection guidelines aimed at maximising heterogeneity and repeatability of the dataset. In particular, the acquisition process was designed to vary the viewpoint of the target with respect to the sonar, both in terms of sensor–target distance and relative orientation, and to include different scene configurations, such as single, target settings with different backgrounds, repeated observations of the same target, and scenes containing different targets. Operationally, the vehicle was manoeuvred around selected viewpoints through controlled horizontal translations, yaw rotations and vertical displacements, while also varying the sonar configuration, the vehicle altitude, and the scenario layout. Additional care was devoted to the recording of FLS configuration parameters during

acquisition and to the synchronisation with other acoustic sensors, so as to reduce interference artefacts and preserve consistency of the collected imagery. These guidelines were adopted to ensure that the resulting FLS datasets were not only sufficiently large, but also structurally diverse and representative of different acquisition conditions.

From a dataset, design perspective, the collection logic was deliberately oriented toward heterogeneity rather than redundancy. The data should represent different scenarios in which the objects appear, with variability in target angle, background settings, and distance from the seabed, and that overly similar or homogeneous patterns should not dominate the dataset. This principle is especially relevant in underwater sonar perception, where repeated acquisition of near, identical samples can easily reduce the effective robustness of the trained model. The collection workflow therefore aimed to expose the future models to a broad range of target presentations and scene conditions already at the data, acquisition stage.

A major trustworthiness, relevant component of the pipeline was the use of data augmentation. The ATR methodological framework already identified augmentation as essential because labelled FLS datasets are difficult and costly to acquire, and because underwater models must generalise beyond the exact observations seen during collection. The design rationale was therefore to strengthen robustness, reduce overfitting, and better simulate the variability encountered in real field conditions. This augmentation logic was concretely instantiated in the actual training pipeline used for the high-frequency ATR model: horizontal flipping, rotation, scaling, translation, HSV-based intensity variation, and mosaic augmentation.

Overall, data collection throughout the initial pilot execution should therefore be understood as a complete supervised-learning data pipeline rather than as raw sensing alone. It included the acquisition of dedicated low-frequency and high-frequency FLS data, the use of a structured four, class target taxonomy (ANVIL, MANNEQUIN, SEWER, and TIRE), the creation of explicit train/validation/test splits, the production of manual polygon-based instance, segmentation annotations, and the integration of the collected data into a concrete augmentation and training workflow. Through this process, the initial pilot established the empirical basis on which the two ATR sonar models were trained, validated and later assessed within the FAITH framework.

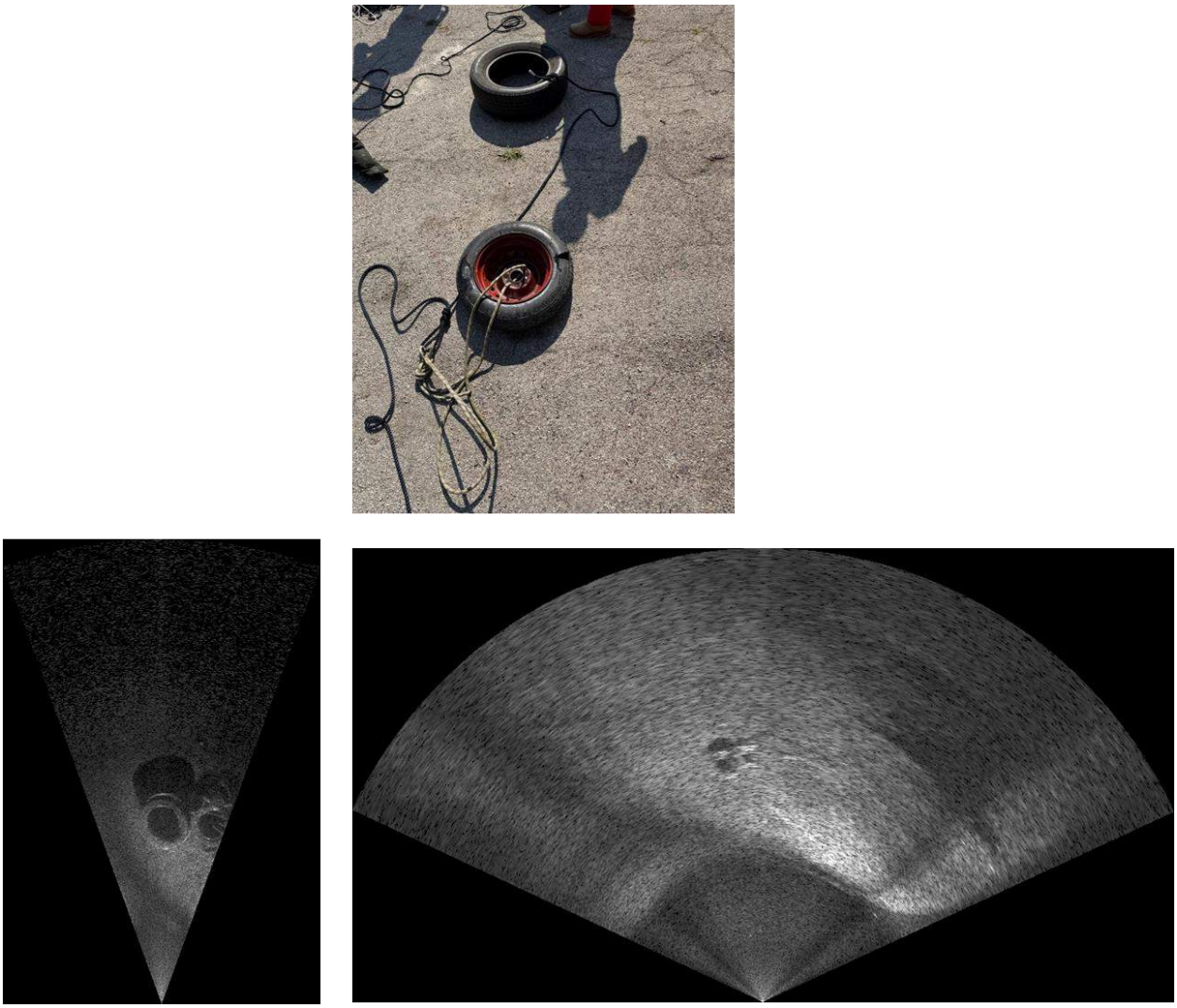


Figure 20: Example of tire target placed on the seabed. On the left, the high-frequency frame; on the right, the low-frequency one.

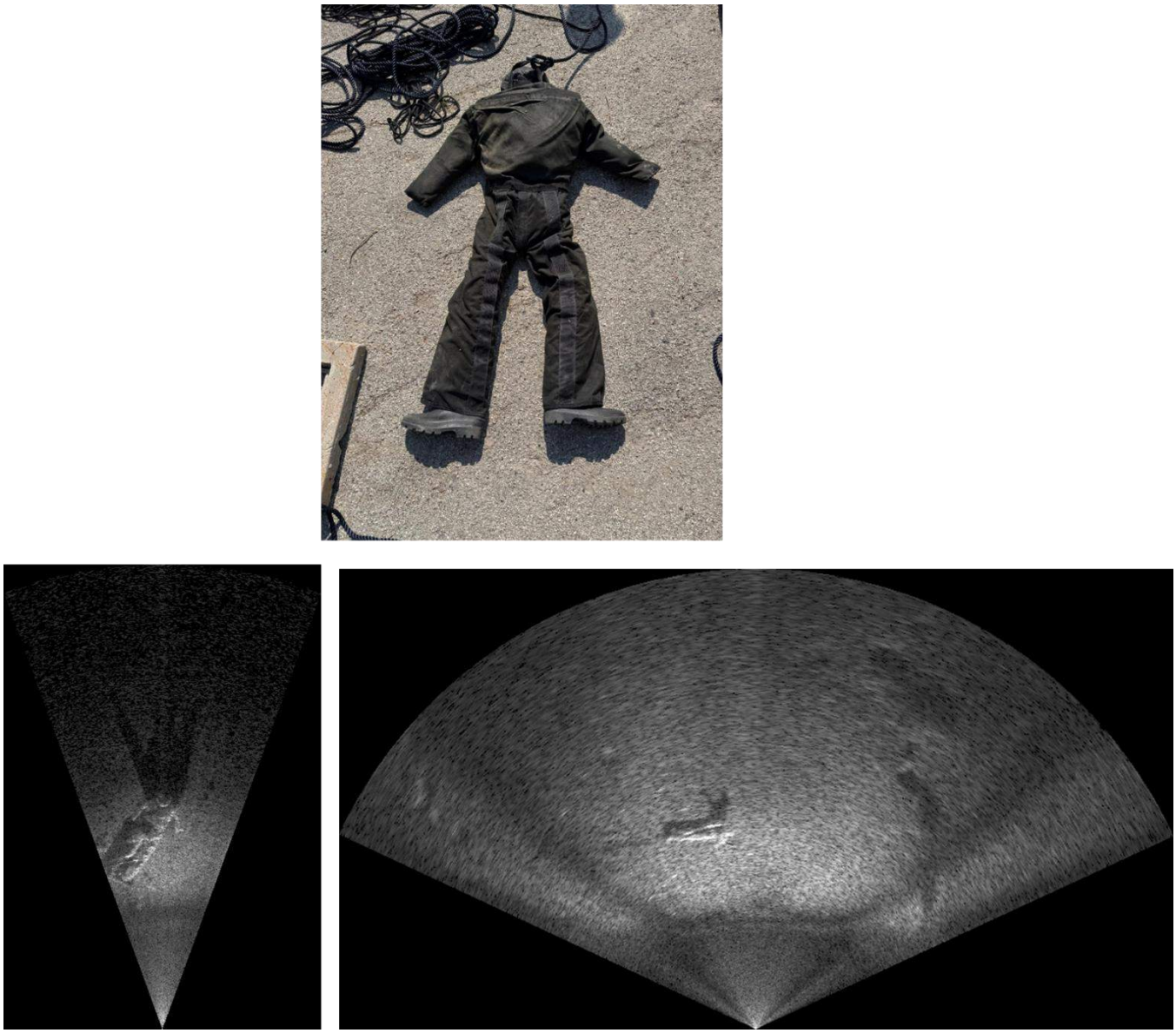


Figure 21: Example of mannequin target placed on the seabed. On the left, the high-frequency frame; on the right, the low-frequency one.

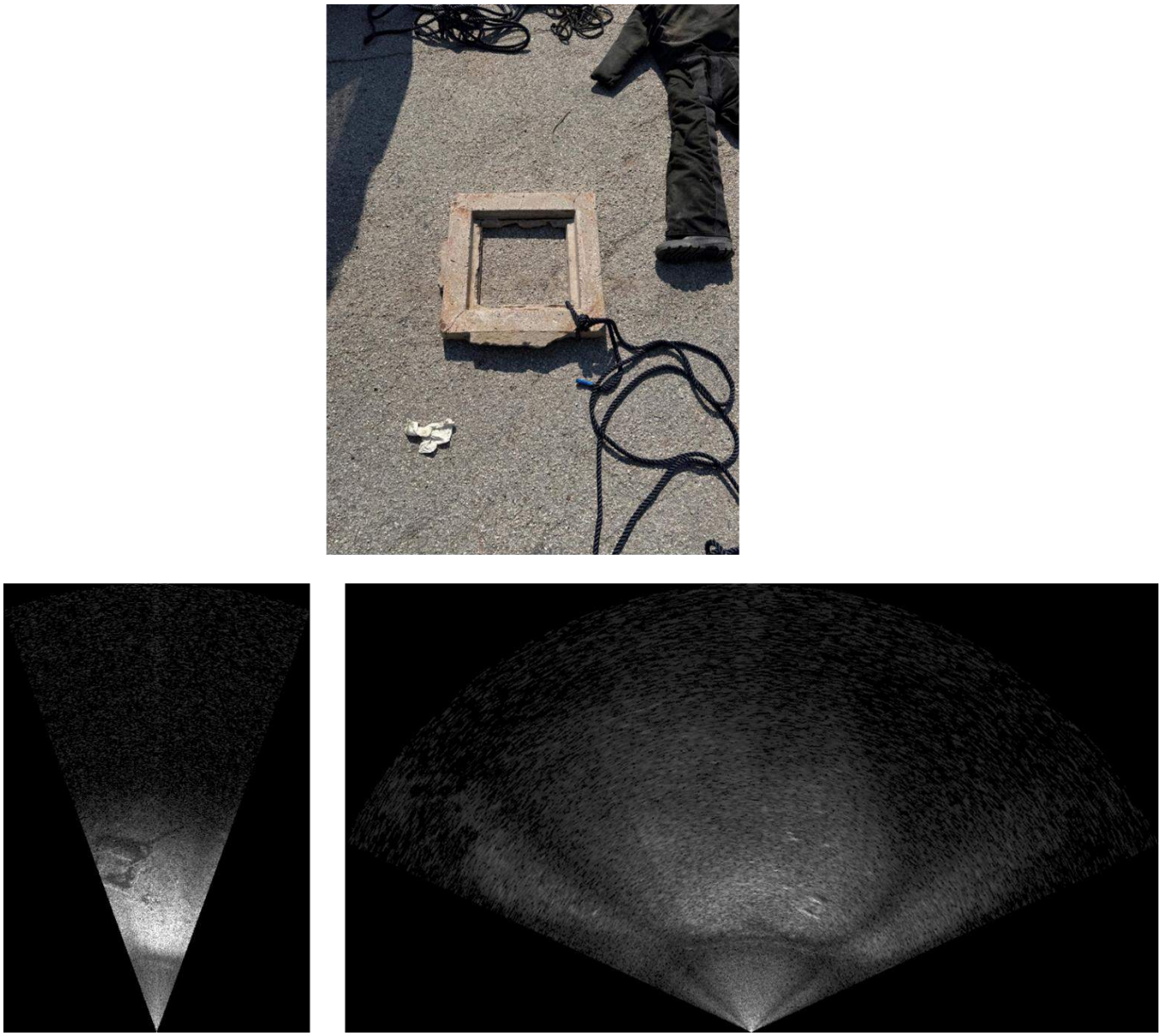


Figure 22: Example of sewer target placed on the seabed. On the left, the high-frequency frame; on the right, the low-frequency one.

2.5. Integration into LSP workflows: Lessons learned

The integration of the ATR and AC systems into the LSP workflows during the initial pilot generated a set of lessons that are relevant not only for the continuation of the Ravenna activities, but also for the preparation of the replication phase. These lessons emerge from the way the pilot was actually executed: through a progressive sequence of exploratory campaigns, structured data collection, refinement of the AI systems, staged end-user training, and first operational use with the APRA and the nautical detachment of the Fire Brigade. Taken together, these activities showed that the successful integration of underwater AI systems into port workflows depends on a balanced combination of operational feasibility, technical maturity, and progressive user involvement.

Operational lessons

A first operational lesson is that the integration of AI-enabled underwater systems into port workflows benefits from a staged deployment strategy, rather than from a direct jump to full operational use. In Ravenna, the pilot moved from an initial contextualisation of sensing limitations, to structured target-based data collection, to testing with trained models, and only afterwards to direct interaction with end users. This progression proved important because it allowed the consortium to align the originally planned use cases with what was actually feasible in the field, while preserving safety, realism and operational credibility. The lesson for the LSP workflow is therefore that underwater AI systems should enter operational routines progressively, with each deployment stage used to reduce uncertainty before exposing the system to more realistic user, facing conditions.

A second operational lesson is that end-user trust and adoption cannot be assumed from technical performance alone. The pilot experience showed that operators need to understand not only what the systems do, but also how they communicate uncertainty, how human intervention is preserved, and what safeguards are visible during use. The combination of the Ravenna stakeholder workshop, field-based practical training, and the later update session before operational use demonstrated that training must be treated as an integral part of workflow integration. In this sense, the pilot confirmed that meaningful operational uptake depends on a human-centred process in which users are first informed, then exposed to the systems in practice, and finally re, aligned when the SUT evolves.

A third operational lesson concerns the importance of scenario realism and logistical adaptability. The pilot showed that some scenarios originally foreseen at planning stage could not be validated convincingly in the actual Ravenna conditions, either because sufficiently evident anomalies were not available or because bathymetric differences were not large enough to support meaningful change, detection tests. The practical response was not to abandon the pilot objective, but to

refine the scenarios into forms that remained operationally meaningful and technically demonstrable. This demonstrated that integration into LSP workflows requires flexibility in matching use cases to real environmental and logistical constraints, including site accessibility, authorisations, target availability, and the practical need to avoid interference with port operations.

Technical lessons

From a technical perspective, one major lesson is that raw AI functionality is not sufficient for operational integration; what matters is the extent to which the AI is embedded in a usable supervisory workflow. For ATR, the initial pilot showed that the system became significantly more useful once it evolved from a detection prototype into a more complete operational application combining perception, mission-state visibility, operator interaction, and explainability/reporting support. For AC, a similar lesson emerged: the planning algorithm became much easier to integrate into the LSP workflow when the operator could explicitly define coverage areas, visualise the generated path, inspect coverage progression, and insert obstacles or risk areas through the interface. In both cases, the lesson is that technical maturity in this domain must include not only algorithmic performance, but also transparency of system behaviour at interface level.

A second technical lesson concerns the need for pipeline refinement driven by field evidence. For ATR, the Ravenna experience made clear that effective port deployment required a more realistic sensing logic than a single generic recognition mode. This is why the pilot converged toward a two, stage sonar workflow, combining a lower, frequency long-range search mode with a higher, frequency close-range inspection mode. At the same time, the pilot confirmed the centrality of dedicated pilot data collection, expert annotation, and augmentation, rich training pipelines in order to cope with the scarcity and variability of underwater sonar data. The lesson here is that robust integration into real workflows requires tight coupling between field campaigns, dataset evolution, retraining, and interface refinement.

A third technical lesson is that human-on-the-loop control is not an accessory feature but a core design requirement for underwater AI systems in port environments. The initial pilot repeatedly showed that user acceptance depends on the possibility to supervise missions, inspect outputs, and intervene when needed. Technically, this translated into alarm acknowledgement, operator takeover logic, explicit mission-state visualisation, risk-aware interaction for ATR, and map-based monitoring and override capability for AC. This means that future workflow integration should continue to treat telemetry visibility, decision traceability, and intervention channels as part of the core system architecture rather than as external add-ons.

Replication pilot implications

The lessons from the initial pilot have direct implications for the replication phase in Valencia. First, the replication pilot should continue to adopt the same iterative validation logic, using the initial pilot not as a finished baseline but as a source of evidence for further system refinement. The current preparation of the Valencia phase already reflects this logic, since it is explicitly oriented toward validating and refining the AI solutions deployed in Ravenna, with a focus on improved trustworthiness, operational reliability and scalability.

Second, the replication pilot should preserve the principle of close alignment between technical scope and site feasibility. The Ravenna pilot experience showed that scenario selection must remain sensitive to what the test site can actually support in terms of logistics, environmental conditions and observable events. For this reason, the original scenarios that could not be robustly demonstrated in the initial pilot should remain under consideration for Valencia, but only as the local environment permissions, accessibility and operational conditions make them meaningful and demonstrable. In this sense, the replication pilot should be understood not as a mechanical repetition of Ravenna, but as a second context in which the validated workflow can be extended selectively and under better, aligned conditions where available.

Third, the replication phase should continue to invest in stakeholder preparation and operator readiness before testing begins. The initial pilot made clear that user engagement is most effective when it combines early awareness, practical exposure and targeted re-training on updated system behaviour. This has a direct implication for the Valencia pilot, where training and capacity, building activities should not be treated as parallel dissemination actions, but as operational prerequisites for meaningful pilot use. The current replication preparation already points in this direction, with end-user recruitment underway, infrastructure finalisation in progress, and training planned ahead of practical deployment.

3. Application of FAITH Tools

3.1. Application of the FAITH AI TAF V1 methodology

In the initial underwater robotics pilot, the FAITH trustworthiness methodology has been applied to AI and human assets in isolation at asset level, rather than to the pilot as a fully connected socio-technical system. The broader in-context modelling of interdependencies between components, users, organisational actors and operational processes is planned for the following phase of the FAITH journey. For LSP4, this methodology has not remained theoretical. It has been operationalised through the actual use of the FAITH tooling environment: TrustGuard for the in-isolation trustworthiness assessment of AI assets, trustSENSE for team-maturity assessment, and the FAITH AI Model Hub for documenting the relevant model and data assets through the

passport logic. The FAITH KB provided the underlying risk-assessment structure used by TrustGuard, and LSP4 also contributed to the KB refinement process by proposing additional underwater-robotics threats and controls that better reflect the port and underwater-survey domain

3.1.1 Cartography

In the FAITH AI_TAF v1, the Cartography phase is the initialization phase in which the boundary of the assessment is explicitly defined. At this stage, the purpose is to identify the AI system under assessment, clarify its intended use, determine the lifecycle stage under evaluation, identify the relevant AI assets and the responsible AI team, and build both an asset model and a user/team model for the assessment. In line with the assumptions of TAF v1, the present LSP4 exercise is therefore conducted in isolation at asset level, meaning that individual AI and human assets are assessed separately, with no modelling yet of cascading risks or of the wider socio-technical interdependencies that characterise the full operational ecosystem. That broader in-context assessment belongs to the next FAITH phase and is planned from M33 onwards through the SSM / TAF-v2 route.

For the underwater robotics pilot, the cartography exercise has been carried out separately for the two AI applications under assessment, ATR and AC, while also accounting for the human teams that design, validate, deploy, supervise, and use them. For ATR, the cartography phase identified as key AI assets the FLS datasets collected during the pilot, the corresponding annotated training, validation and test data, the trained perception models used for sonar-based target recognition, the inference and visualisation software components, and the lifecycle evidence associated with those assets. In parallel, the human side of the cartography identified the UNIFI AI technical team as the owner of the core AI assets, together with the field engineers, testers, deployers, and domain users involved in interpretation and operational supervision, in particular APRA and the Ravenna Fire Brigade. For AC, the cartography phase identified the coverage-planning and replanning algorithm, the multibeam-derived bathymetric validation and test area datasets, the route-generation and mission-supervision software components, the operational logs and interface elements, and the associated human assets responsible for development, deployment and supervision.

The FAITH tools support this phase in a complementary way. TrustGuard is the main tool that operationalises Cartography in the workflow: the risk assessor creates a new assessment, names it, selects the relevant AI assets, chooses the lifecycle stage, and then proceeds to the criticality assessment. In the same phase, TrustGuard also requires the maturity information of the teams interacting with the asset, thereby linking the technical asset boundary with the human-asset boundary. The user can also move from TrustGuard to the FAITH ModelHUB, where the AI assets are documented and reviewed through the relevant metadata records. This is important because,

in FAITH, cartography is not intended as a purely descriptive exercise; it is the phase in which the assessment boundary becomes operational, structured and tool-supported.

On the human side, TrustSENSE supports Cartography by assessing the maturity and readiness of the teams interacting with the AI system under test. TrustSENSE is specifically designed to evaluate the maturity of AI-relevant teams and that its outputs are used by TrustGuard during the Cartography phase. In practical terms, this means that the cartography phase in LSP4 is not limited to identifying models and datasets, but also captures whether the responsible teams have the maturity needed to develop, deploy, supervise and use those assets in a trustworthy way.

Finally, the FAITH ModelHUB, through the Data Passport and AI Model Passport, supports the cartography phase by providing a structured identity layer for AI assets. Indeed, ModelHUB is used as a metadata repository designed to enhance discovery, traceability, interoperability, transparency and accountability for both datasets and AI models. This makes it particularly relevant to Cartography, because it enables the risk assessor to anchor the asset-level assessment to explicit metadata about provenance, intended use, governance and lifecycle information. In the current pilot, this support is already concrete for the ATR branch, where the relevant data and model assets have been documented through the passport logic, while AC assets have not yet been fully integrated into the AI Model Passport layer.

3.1.2 Threat Analysis

After defining the assessment boundary, we carried out the threat-analysis phase by focusing on the concrete threats that could affect the ATR and AC assets in the current underwater robotics pilot. This was not approached as a generic discussion on AI risks, but as a structured activity aimed at identifying the threats that are genuinely plausible in our deployment context, given the underwater environment, the port domain, the role of human supervision, and the current maturity of the systems.

For ATR, the threat analysis concentrated on the threats that can undermine underwater target recognition and trustworthy operator support. In practical terms, this included the risk of misclassification, false positives and false negatives, degradation under difficult sonar conditions, weak robustness outside the validated data distribution, insufficient interpretability of detections, and instability of the broader perception application. These issues matter because an ATR failure in this domain is not just a technical error: it may lead to missed hazardous targets, unnecessary follow-up actions, weaker operator confidence, or unsafe operational assumptions.

For AC, the threat analysis focused on route generation, coverage behaviour and mission execution. The most relevant issues were unsafe or suboptimal routes, reduced generalisability in production settings, degradation under changing external conditions, weak traceability of planning decisions, incomplete coverage, and poor behaviour under environmental variability. In

the port environment, such threats may affect not only mission efficiency, but also the safety and usefulness of the bathymetric mission itself.

The main tool used in this phase was TrustGuard (for ATR), which required us to estimate the frequency of occurrence of a broad set of threats for each asset. This was valuable because it forced us to distinguish between threats that are theoretically possible and threats that are actually material for the current underwater robotics context. In our case, the exercise made clear that the most relevant threats are mostly linked to reliability, environmental variability, weak safeguards, malfunction, and insufficient production robustness, while many classical external-attack or disclosure-oriented threats remain secondary in the present pilot profile.

A brief but important point concerns the FAITH Knowledge Base. For the purpose of the pilot we did not only use the threat inventory already available through the FAITH tooling environment, but also contributed to its refinement by proposing additional underwater-robotics-specific threats through the KB-enhancement activity. This included, for example, threats such as misclassification of structural defects and objects, unsafe route near sensitive structures, tampered mission parameters, opaque coverage decisions, and inspection neglect due to biased path sampling. This contribution was useful because it made the threat base more representative of the concrete trustworthiness issues encountered in underwater inspection and autonomous surveying.

3.1.3 Impact Assessment

Once the most relevant threats had been identified, we carried out the impact-assessment phase in order to understand how severe their consequences would be for the assessed assets. In practice, this phase was used to translate the previously identified threats into concrete trustworthiness consequences for the pilot, rather than to leave them as abstract technical concerns. The main purpose was to understand how much each threat could affect the trustworthiness properties that matter most in underwater-port operations.

This work was performed mainly through TrustGuard (for ATR, qualitatively for AC), which required us to evaluate the consequence level associated with the compromise of the main trustworthiness properties, including safe operation, security and resilience, interpretability and explainability, privacy, fairness, transparency, and accuracy/robustness. This was especially useful because it forced us to reason in operational terms. Instead of stopping at statements such as “the model may fail” or “the planner may degrade,” we had to judge what such failures would actually mean in terms of mission effectiveness, safety, usability, auditability and operator trust.

For ATR, this meant assessing the consequences of missed detections, spurious detections, unstable behaviour outside validated conditions, and insufficiently understandable outputs. In our pilot context, these issues translate into consequences such as missed hazardous objects,

unnecessary inspection interruption, delayed intervention, and reduced confidence in the AI support layer. As a result, the most material consequence areas for ATR clearly emerged as Validity and Reliability and Safety, followed by Explainability and Interpretability and Accountability and Transparency.

For AC, the impact exercise focused on what would happen if the coverage planner or the bathymetric mission-support chain became unreliable, unsafe or insufficiently transparent. Here the main consequences concerned incomplete area coverage, reduced usefulness of the resulting bathymetry, inefficient mission execution, and potential unsafe route behaviour near infrastructure or obstacles. Again, the strongest consequence dimensions were not evenly distributed across all trustworthiness properties: they were mainly concentrated on Validity and Reliability, Safety, and Security and Resilience, while privacy-related consequences remained lower in the present pilot conditions.

What we did in this phase was therefore to clarify which trustworthiness consequences actually matter most for the underwater robotics pilot. This helped us avoid treating all trustworthiness dimensions as equally critical and created a much more useful foundation for the later vulnerability and risk steps.

3.1.4 Vulnerability Assessment

After the threat and consequence phases, we carried out the vulnerability assessment in order to identify the concrete weaknesses that could make the assessed assets more susceptible to the previously identified threats. In the pilot, this was one of the most practically useful steps, because it moved the exercise from abstract reasoning to a much more operational review of the safeguards, lifecycle practices and technical weak points of the assets under assessment.

In practice, this phase was again implemented mainly through TrustGuard (for ATR, qualitatively for AC), which presented us with tailored control questions depending on the asset and lifecycle stage. For each asset, we had to answer a structured set of questions on whether relevant protections, controls and engineering practices were already in place. This meant that the exercise was not simply about “listing weaknesses,” but about checking whether the kinds of safeguards expected around the asset actually existed in the current pilot.

For the ATR model, the vulnerability assessment concentrated on the robustness and governance of the deployed perception asset. The strongest areas were related to data control, internal management of the training pipeline, use of trusted data, and availability of performance indicators for correct functioning. At the same time, the assessment also made visible weaker areas around security governance, access and identity management, systematic vulnerability monitoring, and some lifecycle controls linked to disclosure, drift, and formalised security processes. Beyond governance, the vulnerability profile also reflected technical exposure to

turbidity, sonar artefacts, variable environmental conditions, and underrepresented scenarios in the training data.

For the ATR dataset, the vulnerability assessment shifted attention toward dataset governance, provenance, representativeness, protection and suitability for training and validation. Here the key questions concerned whether the data were controlled through trusted pipelines, whether quality checks were in place, whether diversity was sufficient to support generalisation, and whether there were governance measures around access and provenance. This was especially important because a large part of the residual ATR risk is tied to whether the dataset fully represents the conditions in which the model is later expected to operate.

For the AC asset, the vulnerability profile was different and more strongly linked to the end-to-end autonomy chain. The key weaknesses concerned sensor calibration, robustness of bathymetric sensing, integration between planning and execution, behaviour under changing seabed conditions, support for obstacle handling, and the availability of sufficient logging and decision provenance. In addition, there are also human-side vulnerabilities if operators are not fully prepared to interpret route evolution, assess coverage behaviour, or intervene appropriately when the mission becomes uncertain. This made AC a good example of an asset whose vulnerability profile is simultaneously technical, operational and organisational.

A relevant complementary contribution came from trustSENSE, because the maturity of the teams interacting with the AI assets influences how some vulnerabilities materialise in practice. Stronger team maturity reduces the risk that a technical issue propagates unchecked into operational harm, while weaker organisational maturity can amplify the real-world effect of a technical weakness. In this way, the human side of vulnerability was not ignored, but integrated into the overall understanding of the pilot's trustworthiness posture.

3.1.5 Risk Analysis

After completing the threat, impact and vulnerability phases, we carried out the risk-analysis phase in order to identify which issues should be treated as the most important priorities for the current pilot assets. In practice, this phase allowed us to move from a set of separate findings to a more synthetic judgement on where the most material trustworthiness risks actually lie.

For ATR, the risk analysis showed that the most important risks are those where high-impact threats intersect with high operational and technical vulnerability. In our case, this means above all the risk of missed detections or spurious alarms in safety-relevant inspection scenarios, reduced robustness outside the validated data distribution, and insufficiently interpretable outputs when the AI result has operational implications. The exercise showed clearly that ATR risk is not defined only by average performance values, but by whether the perception chain remains dependable under real field conditions and whether the operator is given enough evidence to

trust and supervise the output appropriately. As a result, the most relevant ATR risks are concentrated around Validity and Reliability, Safety, and, at a second level, Explainability and Interpretability together with Accountability and Transparency.

For AC, the dominant risks are linked to route-planning failure, incomplete coverage, insufficient obstacle handling, sensor-related inaccuracies, and weak traceability of planning decisions. In other words, the main concern is not that the planner cannot produce a path, but that the resulting autonomous behaviour may become less dependable when exposed to real deployment variability, especially under changing external conditions or with still-prototypal system integration. This is coherent with the field evidence gathered during the pilot, where the AC logic already showed strong functional validity, but the full system still displayed residual fragilities linked to platform and integration maturity.

Here too, TrustGuard was the main tool supporting the exercise (for ATR), because it helped combine the previous phases into a risk-oriented picture rather than leaving them as isolated results. This is what allowed us to distinguish between risks that are theoretically present but practically marginal, and those that really need mitigation in the present pilot cycle. trustSENSE remained relevant as well, because the maturity of the responsible teams affects how some risks are amplified or contained in practice.

What we actually did with the output of this phase was to use it as a basis for prioritising engineering and deployment effort. The phase was therefore not only descriptive: it directly supported the identification of where improvements should be concentrated in the next cycle.

3.1.6 Countermeasures

The last phase of the current FAITH cycle was the definition of countermeasures to reduce the most relevant risks identified for the assessed assets. In the pilot, this phase was not approached as a theoretical list of possible controls, but as a practical step in which the previous assessment work was translated into concrete technical and organisational actions. The goal was to identify which measures could realistically improve the SUT and strengthen the conditions under which users can supervise and trust them appropriately.

For ATR, the countermeasure strategy focuses first on strengthening data and model robustness. The measures already implemented or planned include expansion and diversification of the training dataset, additional field trials under both clear and turbid water conditions, further fine-tuning and maintenance of the CNN models, and cross-validation against representative port scenarios. These actions are intended to reduce the current exposure to degraded performance outside the initial validation envelope and to improve the stability of the perception system under realistic environmental variability.

To strengthen Safety, the ATR branch introduces an out-of-distribution detection layer so that live frames can be compared against the training distribution and graded responses can be triggered when uncertainty or abnormality is detected. In practice, this adds a runtime trustworthiness mechanism that helps prevent unsafe overreliance on nominal model outputs when the input conditions deviate from the validated data domain. To strengthen Explainability and Interpretability, the pilot uses Grad-CAM heatmaps together with retrieval of similar training examples, so that operators can better understand which visual evidence drove each detection and can compare the current observation with known examples from the training set. This makes the ATR output more understandable and more reviewable in operational use. To strengthen Accountability and Transparency, the system relies on more complete logging of sensor states, model and dataset versions, detections, alerts, and configurations, together with operator override capabilities and real-time visual monitoring of detections. These elements are particularly important because they support auditability, post-mission reconstruction of events, and more disciplined supervision of the perception chain during deployment.

For AC, the main countermeasures are centred on deployment robustness and mission-level dependability. These include rigorous sensor calibration, robust bathymetric data processing, and extensive simulation and field trials aimed at checking the behaviour of the AC chain under a broader range of operational conditions. In parallel, the planning architecture itself acts as a mitigation measure through a two-tier structure, in which a global RRT*-style coverage planner is complemented by a local planner for obstacle avoidance and short-horizon adaptation. This architecture is intended to make the system more robust to local disturbances and to reduce the risk of unsafe or inefficient route execution. The pilot also includes dynamic route recalculation, detection and avoidance of unexpected obstacles, and manual override capability among the core mitigation mechanisms. These are direct controls for Safety and human oversight, because they ensure that the planner is not treated as an immutable black box and that the operator can intervene whenever route execution becomes uncertain or inconsistent with mission expectations. To strengthen Explainability and Transparency, AC relies on detailed logs, decision trails, real-time visualisation of routes and coverage evolution, and post-mission analyses comparing chosen and unchosen paths. These measures are essential because, in AC, interpretability depends less on explaining an output label and more on allowing operators to understand why a certain path was selected, how coverage evolved over time, and how the planner behaved in relation to the mission constraints. For Fairness, the pilot foresees auditing of survey effort across all requested regions, including more complex or high-risk port zones, so that the system does not systematically privilege only easier survey areas. In this context, fairness is not interpreted in social terms, but rather as balanced and non-distortive operational allocation of survey attention across the relevant mission space.

The FAITH tools supported this phase in complementary ways. TrustGuard provided the main structure for the countermeasure logic by connecting identified risks to candidate controls. trustSENSE also remained relevant because some mitigation measures concern the human teams rather than the AI assets alone, especially where supervision, interpretation of outputs, and response to anomalies remain part of the trust model. For ATR, the passport-based governance logic of the Model Hub also supports mitigation by making the identity, intended use and operational boundaries of the assets more explicit, thereby reducing the risk of inappropriate reuse or overclaiming of readiness. For AC, even though the integration with the FAITH tools has not yet been completed, the same governance logic will help keep the mitigation plan realistic.

3.2 Tool Validation Planning & Training Processes

3.2.1 Dry-run

ATR

The dry-run activity constituted the first structured application of the FAITH AI_TAF v1 within the LSP4 underwater robotics pilot. The objective of this activity was twofold: on the one hand, to validate the applicability of the FAITH methodology in the specific domain of AI-driven underwater inspection; on the other hand, to obtain an initial, evidence-based assessment of trustworthiness risks associated with the ATR system developed by UNIFI.

The dry-run followed the logical sequence of the FAITH AI_TAF, starting from the cartography phase and progressing through criticality, threat frequency, consequence evaluation and vulnerability analysis. In the cartography phase, the ATR system was defined as the system under test within the underwater robotics domain, with a focus on the design lifecycle stage. The system was decomposed into its core AI assets, including three scenario-specific models (M1, M2, M3) and their corresponding training and testing datasets (Dtr1–Dtr3 and Dte1–Dte3). These scenarios cover detection of sunken objects or subsea intruders, inspection of underwater infrastructures, and on, demand ship hull inspection.

The criticality assessment revealed that, although the ATR system does not formally fall under the high-risk categories defined in Annex III of the AI Act, it operates in a context characterised by high operational criticality. The system is deployed within port environments and supports activities conducted by port authorities, which qualify as essential entities in the water transport sector. Moreover, failures in the ATR system may lead to safety, critical consequences, such as missed detection of hazardous objects or incorrect triggering of countermeasures. Based on these considerations, the system was classified as high-criticality from an operational perspective.

	ANSWER	Comment		
AI SYSTEM NAME	AUTOMATIC TARGET RECOGNITION - ATR	We have 3 possible scenarios for underwater ATR with relatives models and datasets: Detection of sunken objects and/or subsea intruder (M1, Dtr1, Dte1), Inspection of Underwater Infrastructures (M2, Dtr2, Dte2), On-demand ship hull inspection (M3, Dtr3, Dte3)		
SECTOR	UNDERWATER ROBOTICS			
LIFECYCLE STAGE	DESIGN			
AVG score of all stakeholders' technical proficiency (ONLY IF LIFECYCLE STAGE=DEPLOY) (please visit D2.1 for this)				
AI ASSET TYPE*	**AI ASSET NAME**	**LIFECYCLE STAGE**	**TEAM OWNER per AI Asset (only if lifecycle stage= DESIGN/DEVELOP)**	**AI MATURITY FOR TEAM (please visit D2.1 for this)**
AI ASSET - Models *(insert below more lines if necessary)*	M1	DESIGN	UNIFI AI	Very High
AI ASSET - Models (insert below more lines if necessary)	M2	DESIGN	UNIFI AI	Very High
AI ASSET - Models (insert below more lines if necessary)	M3	DESIGN	UNIFI AI	Very High
AI ASSET - Datasets *(insert more lines if necessary)*	Dtr1	DESIGN	UNIFI FIELD ENG	High
AI ASSET - Datasets *(insert more lines if necessary)*	Dtr2	DESIGN	UNIFI FIELD ENG	High
AI ASSET - Datasets *(insert more lines if necessary)*	Dtr3	DESIGN	UNIFI FIELD ENG	High
AI ASSET - Datasets *(insert more lines if necessary)*	Dte1	DESIGN	UNIFI FIELD ENG	High
AI ASSET - Datasets *(insert more lines if necessary)*	Dte2	DESIGN	UNIFI FIELD ENG	High
AI ASSET - Datasets *(insert more lines if necessary)*	Dte3	DESIGN	UNIFI FIELD ENG	High

Figure 23: ATR Dry-run cartography

The threat assessment phase focused on estimating the frequency of relevant threats in the underwater robotics domain. The results indicate that classical adversarial machine learning attacks, such as poisoning, oracle attacks or white-box adversarial example generation, are considered highly unlikely in this context due to the isolated and non-quarriable nature of the system, which operates onboard AUVs without external exposure. In contrast, higher likelihood is associated with operational and environmental risks, such as system failures, hardware malfunctions, inconsistent sensor data, or human error. This confirms that the ATR system is more exposed to physical, world uncertainties and system-level failures than to typical cyber, attacks observed in internet, facing AI systems.

The consequence assessment highlighted a clear prioritisation of trustworthiness dimensions. Safety, security and resilience, and validity and reliability were assessed as having very high impact, reflecting the potential consequences of incorrect detections or system failures in port operations. Explainability, interpretability and transparency were evaluated as moderate, as they primarily affect trust, usability and root, cause analysis rather than immediate safety. Privacy was assessed as low, given that the system mainly processes underwater infrastructure data and rarely involves personal information. Fairness-related risks were considered negligible, as the application domain involves inanimate objects rather than human subjects or protected groups.

This prioritisation is fully consistent with the characteristics of the underwater robotics domain and with the trustworthiness requirements identified for the pilot.

CRITICALITY ASSESMENT		
Questions	**Useful resources**	**Answer**
Does the AI system falls under the qualification of a high risk AI-system as established by Article 6 of the AI Act, such as the high risks systems listed in AI Act Annex III If the AI system poses a very high risk of harm to the health, safety or fundamental rights of natural persons, including by not materially influencing the outcome of decision making.	Annex III	No
Does the AI system pose a high risk of harm to the health, safety or fundamental rights of natural persons, including by not materially influencing the outcome of decision making? and /or Is the AI system is used by an Operator of Essential services (OES) and the AI system is used for the provision of an essential service as defined in Annex I of NIS 2 Directive?	Annex I of NIS 2 Directive	Yes
Does the AI system pose substantial risk of harm to the health, safety or fundamental rights of natural persons, including influencing the outcome of decision making If the AI system is used by an Operator of Essential services (OES) and the AI system is used for the provision of an important service as defined in Annex 2 of NIS2?	Annex 2 of NIS2	No
Does the AI system pose medium risk of harm to the health, safety or fundamental rights of natural persons, including influencing the outcome of decision making? Is the AI system for the provision of an important service as defined in Annex 2 of NIS2?		No
Does the AI system pose low or no risk of harm to the health, safety or fundamental rights of natural persons, including influencing the outcome of decision making?		No

Figure 24: ATR Dry-run criticality assessment

The vulnerability analysis was performed through a detailed set of control questions targeting the design phase of the ATR system. The results show a high level of maturity in aspects related to data quality and model development. In particular, training datasets are acquired and labelled internally by expert underwater robotics engineers, with multi, level validation procedures. The model design leverages fine, tuning of pre-trained architectures, combined with data augmentation techniques to improve robustness under challenging underwater conditions. Performance indicators such as instance segmentation metrics and inference constraints on onboard hardware are clearly defined and monitored. These elements indicate strong technical robustness at the level of model development and dataset management.

At the same time, the assessment identified several gaps in terms of security governance and lifecycle management. Controls related to access control, identity management, formal security processes, vulnerability monitoring, and integration with organisational security policies are only partially implemented or currently absent. Similarly, advanced privacy-preserving techniques and

structured risk analysis processes for disclosure scenarios have not yet been incorporated. These findings suggest that, while the ATR system demonstrates high technical maturity, particularly in perception and data handling, additional effort is required to strengthen its security, by, design and governance aspects.

Threat	Threat Definition *(from: https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms)*	Question	Answer (use the "Default" value if you do not know the answer)	Comment FVP
Label modification	An attack in which the attacker corrupts the labels of training data. This sub-threat is specific to Supervised Learning.	**How often do attackers in similar AI systems modify the training data in this sector?**	Never	Modifying training data is typically a concern in supervised learning scenarios where attackers have access to the training dataset. In your case, with automated underwater drones, this threat is less likely as the training data is not exposed to external manipulation.
Evasion	A type of attack in which the attacker works on the ML algorithm's inputs to find small perturbations leading to large modification of its outputs (e.g. decision errors). It is as if the attacker created an optical illusion for the algorithm. Such modified inputs are often called adversarial examples. Example: the projection of images on a house could lead the algorithm of an autonomous car to take the decision to suddenly make it brake	**How often do evasion attacks occur in similar AI systems in this sector?**	Once every 10 years	Evasion attacks are more common in adversarial environments where attackers can manipulate inputs. For underwater drones, such attacks are rare but possible, especially if an adversary gains access to the system.
Use of adversarial examples crafted in white or grey box conditions	In some cases, the attacker has access to information (model, model parameters, etc.) that can allow him to directly build adversarial examples. One example is to directly use the model's gradient to find the best perturbation to add to the input data to evade the model.	**How often do attacks with adversarial examples crafted in white or grey box conditions occur in this sector?**	Never	Adversarial examples are less common in physical systems like underwater drones, especially in white or grey box conditions, as they require detailed knowledge of the system.
Oracle	A type of attack in which the attacker explores a model by providing a series of carefully crafted inputs and observing outputs. These attacks can be previous steps to more harmful types, evasion or poisoning for example. It is as if the attacker made the model talk to then better compromise it or to obtain information204 about it (e.g. model extraction) or its training data (e.g. membership inferences attacks and Inversion attacks). Example: an attacker studies the set of input output pairs and uses the results to retrieve training data.	**How often do oracle attacks occur in similar AI systems in this sector?**	Never	Oracle attacks involve querying the model to gain information. This is less applicable to underwater drones operating in isolated environments.
Failure or malfunction of ML application	This threat refers to ML application failure (e.g. denial of service due to bad input, unavailability due to a handling error). Example: the service level of the support infrastructure of the ML application hosted by a third party is too low compared to the business needs, the application is regularly unavailable. Note that this threat does not consider failure of business use cases (for example, the algorithm fails because it is not accurate enough to handle all real-life situations it is exposed to).	**How often do such ML applications fail or malfunction in this sector?**	Once a year	Failures or malfunctions can occur due to environmental factors, hardware issues, or software bugs, making this a more frequent concern.
Denial of service due to inconsistent data or a sponge example	ML algorithms usually consider input data in a defined format to make their predictions. Thus, a denial of service could be caused by input data whose format is inappropriate. It may also happen that a malicious user of the model constructs an input data (a sponge example) specifically designed to increase the computation time of the model and thus potentially cause a denial of service	**How often does denial of service occur due to inconsistent data or a sponge example in similar AI systems in this sector?**	Once every 2 years	Denial of service due to inconsistent data is a concern, especially if the drones rely on real-time data processing. Sponge examples are less common but could be engineered to disrupt operations.
Poisoning	A type of attack in which the attacker altered data or model to modify the ML algorithm's behavior in a chosen direction (e.g. to sabotage its results, to insert a backdoor). It is as if the attacker conditioned the algorithm according to its motivations. Such attacks are also called causative attacks. Example: massively indicating to an image recognition algorithm that images of dogs are indeed cats to lead it to interpret it this way.	**How often do poisoning attacks occur in similar AI systems in this sector?**	Never	Poisoning attacks involve corrupting the training data, which is less likely in a controlled environment.
Model or data disclosure	This threat refers to the possibility of leakage of all or partial information about the model.12 Example: the outputs of a ML algorithm are so verbose that they give information about its configuration (or leakage of sensitive data)	**How often do the models used in this sector disclose information about their configuration?**	Once every 10 years	
Data disclosure	This threat refers to a leak of data manipulated by ML algorithms. This data leakage can be explained by an inadequate access control, a handling error of the project team or simply because sometimes the entity that owns the model and the entity that owns the data are distinct. To train the model, it is often necessary for the data to be accessed by the model provider. This involve sharing the data and thus share sensitive data with a third party.	**How often are the data manipulated by the ML algorithms leaked in this sector?**	Once every 10 years	
Model disclosure	This threat refers to a leak of the internals (i.e. parameter values) of the ML model. This model leakage could occur because of human error or contraction with a third party with a too low security level.	**How often are the model internals (i.e., parameter values) of the AI system leaked in this sector?**	Once every 5 years	Model internals are usually well-protected in proprietary systems, making leaks less likely.
Compromise of ML application components	This threat refers to the compromise of a component or developing tool of the ML application. Example: compromise of one of the open-source libraries used by the developers to implement the ML algorithm.	**How often are the components of similar ML applications compromised in this sector?**	Once a year	Component compromise is rare but possible, especially if third-party components are used.
Human error	The different stakeholders of the model can make mistakes that result in a failure or malfunction of ML application. For example, due to lack of documentation, they may use the application in use-cases not initially foreseen.	**How often do such ML applications fail or malfunction due to end-user error in this sector?**	Once a year	End-user errors can occur, especially in complex systems like underwater drones, leading to occasional failures or malfunctions.
Cybersecurity incident not reported to incident response teams	This threat refers to the possibility that a project team may not report security incidents to dedicated teams while a policy of mandatory incident reporting has been defined.	**How often do cybersecurity incidents occurring in similar AI systems miss to be reported in this sector?**	Once every 2 years	Underreporting of cybersecurity incidents is a known issue, and while efforts are made to report them, some incidents may still go unreported.

Figure 25: ATR Dry-run Frequency of threats

Trustworthiness characteristic	Question	Answer	Please provide in detail potential business and legal consequences UNIFI
Safety	What is the business consequence level if the **safe** operation of the AI system is compromised?	Very High	A false negative (missed explosive device / structural crack) or a false positive that closes gates or pumps at the wrong moment can injure divers, crew or dock workers and damage high-value vessels.
Security and resilience	What is the business consequence level if the **security or the resilience** of the AI system to adverse events is compromised?	Very High	A cyber-takeover of the ATR or its comms link could be used to sabotage port operations or mask hostile activities → immediate safety & economic impact, regulatory breach
Explainability and interpretability	What is the business consequence level if the AI system is not **interpretable**?	Moderate	Engineers need to understand why alarms or triggers are flagged to plan countermeasures; lack of interpretability slows root-cause analysis but seldom causes direct harm.
	What is the business consequence level if the AI system is not **explainable**?	Moderate	Same rationale as interpretability; the business impact is delays and extra cost, not life-threatening situations.
Privacy	What is the business consequence level caused by a **privacy** compromise of the AI system?	Low	AUV targets are of the same nature of quay walls, ship hulls or seabed; personal data are rarely captured. Exposure mainly reveals sensitive infrastructure layouts, not individual identities.
Fairness with harmful bias managed	What is the business consequence level if the AI system is not **fair**?	Negligible	The model identifies inanimate underwater objects; no protected classes affected, so unfairness would have negligible legal/business impact.
Accountability and Tranparency	What is the business consequence level if the AI system is not **transparent**?	Moderate	Poor transparency increases legal exposure but rarely causes immediate operational loss.
Validity and reliability	What is the business consequence level if the AI system is not **accurate**?	Very High	Accuracy is crucial: systematic mis-detections can mask serious hull or quay damage, leading to costly emergency repairs or shutdowns.
	What is the business consequence level if the AI system is not **robust**?	Very High	

Figure 26: ATR Dry-run Business Consequence Level

Control	Relevant AI Asset for the control question (Replace with the AI Assets of your SUT)	Template Control Questions (Please replace "Dtr", "M1", "M2" with the names of the AI assets declared in the cartography phase. Add more/remove lines if needed)	Answer	Comment (please add here your comments about the questions)
Ensure reliable sources are used	Dtr1	*In the context of Dtr, have you ensured that reliable sources were used to label the data?*	Yes	Labels created in-house by expert underwater robotics engineers with a minimum 2-person quality check
Choose and define a more resilient model design	M1	*In the context of M1, have you chosen and defined a resilient model design?*	Yes	learned features to enhance the performance of instance segmentation tasks in a context with limited and highly specific
Assess the exposure level of the model used	M1	*In the context of M1, have you assessed the exposure level of the model to poisoning adversarial attacks?*	No	
Enlarge the training dataset (by adding more diverse data)	M1	*n the context of M1, have you enlarged the training dataset by adding diverse data to increase its robustness to poisoning attacks?*	Yes	Data augmentation to expand the dataset dimensionality
Apply a RBAC model, respecting the least privileged principle	M1	*In the context of M1, have you applied an RBAC model, respecting the least privileged principle?*	No	
	Dtr1	*In the context of Tr1, have you applied an RBAC model, respecting the least privileged principle?*	No	
Ensure ML applications comply with data security requirements	M1	*In the context of M1, have you ensured that the model complies with data security requirements?*	No	
Define and monitor indicators for proper functioning of the model	M1	*In the context of M1, have there been defined indicators of proper functioning of the model?*	Yes	Standard instance segmentation metrics (mAP, etc.) along with the required inference time on the underwater robot hardware
Integrate ML specificities to awareness strategy and ensure all ML stakeholders are receiving it	M1	*In the context of M1, have you integrated ML specifities to awareness strategies and ensured all ML stakeholders are receiving it?*	No	
Control all data used by the ML model	M1	*In the context of M1, have you controlled all data used by the model?*	Yes	Data controlled by expert underwater robotics engineers with a minimum 2-person quality check
Ensure reliable sources are used	Dtr1	*Is Dtr reliable?*	Yes	Data acquired directly from the underwater robot from expert underwater engineers
	M1	*Is the base model used for M1 reliable?*	Yes	Pre-trained weights pulled directly from Ultralytics
Use methods to clean the training dataset from adversarial samples	Dtr1	*Does Dtr include adversarial samples?*	No	
Ensure ML applications comply with identity management, authentication, and access control policies	M1	*In the context of M1, have you ensured that the model complies with with identity management, authentication, and access control policies ?*	No	
Conduct a risk analysis of the ML application	M1	*In the context of M1, have you conducted risk analysis to identify all disclosure scenarios?*	No	
Implement processes to maintain security levels of ML components over time	M1	*Have you implemented processes to maintain security levels of M1 components over time?*	No	
Ensure that models respect differential privacy.	M1	*Does the M1 model respect differential privacy?*	No	
Reduce the information given by the model	M1	*Have you designed the M1 model in such a way that will provide only the necessary information and will not disclose any private information?*	No	
Use federated learning to minimise the risk of data breaches	M1	*In the context of M1, have you used ferated learning to minimise the risk of data breaches?*	No	
Limit the available information about the model architecture	M1	*In the context of M1, have you limited the available information about the model architecture?*	Yes	The model architecture is shared but fine-tuned weights kept internal
Reduce the output information given by the model	M1	*In the context of M1, have the output information given by the model be*	Yes	The model output (e.g., bounding-boxes, confidences, etc.) are translated into user-oriented informations (e.g., risk levels for different object types based on their potential threat)
Ensure ML projects follow the global process for integrating security into projects	M1	*In the context of M1, have it been ensured that the model has followed the global process for integrating security?*	No	
Check the vulnerabilities of the components used so that they have an appropriate security level	M1	*In the context of M1, have you checked the vulnerabilities of the components used so that they have an appropriate security level?*	No	
Integrate ML specificities to existing security policies	M1	*In the context of M1, have there been integrated model specifities to existing security policies?*	No	
Ensure ML applications comply with security policies	M1	*In the context of M1, have you ensured that the model complies with the relevant security policies?*	No	
Include ML applications into asset management processes	M1	*Have you included the M1 model into the asset management processes?*	No	
Ensure ML applications comply with third parties' security requirements		*Have you ensured that the M1 model complies with third parties' security requirements?*		

Figure 27: ATR Dry-run Control Questions

AC

The dry-run activity for the AC component of the LSP4 underwater robotics pilot represented the first structured application of the FAITH AI_TAF v1 to the mission planning and coverage optimisation functionality deployed on AUVs.

Similarly to the ATR case, the activity was conducted as a collaborative desk-based assessment involving technical partners, pilot stakeholders, and domain experts. The cartography phase defined the system under assessment as the AC module within the underwater robotics domain, focusing on the design lifecycle stage. The system was decomposed into its core AI assets, including the main coverage algorithm (Alg1), a validation area dataset (Dval1), and two testing area datasets (Dte1 and Dte2). The validation dataset is used to tune algorithm parameters, while the testing datasets represent unseen real-world areas where the algorithm is evaluated through bathymetry reconstruction tasks.

	ANSWER	Comment		
AI SYSTEM NAME	AUTONOMOUS COVERAGE - AC	We have one algorithm Alg1, one validation dataset Dval1 storing areas upon which the Alg1 will be validated and two testing area datasets Dte1, Dte2 representing areas of real testing scenarios to realize bathymetries upon. The validation dataset Dval1 contains areas used to tune the parameters of the algorithm; the test datasets Dte1 and Dte2 contain unseen areas where Alg1 will be tested and bathymetries realized.		
SECTOR	UNDERWATER ROBOTICS			
LIFECYCLE STAGE	DESIGN			
AVG score of all stakeholders' technical proficiency (ONLY IF LIFECYCLE STAGE=DEPLOY) (please visit D2.1 for this)				
AI ASSET TYPE*	**AI ASSET NAME**	**LIFECYCLE STAGE**	**TEAM OWNER per AI Asset (only if lifecycle stage= DESIGN/DEVELOP)**	**AI MATURITY FOR TEAM (please visit D2.1 for this)**
AI ASSET -Models *(insert below more lines if necessary)*	Alg1	DESIGN	UNIFI AI	Very High
AI ASSET -Models (insert below more lines if necessary)	Dval1	DESIGN	UNIFI FIELD ENG	Very High
AI ASSET -Datasets *(insert more lines if necessary)*	Dte1	DESIGN	UNIFI FIELD ENG	Very High
AI ASSET -Datasets *(insert more lines if necessary)*	Dte2	DESIGN	UNIFI FIELD ENG	Very High

Figure 28: AC Dry-run Cartography

Questions	Useful resources	Answer
Does the AI system falls under the qualification of a high risk AI-system as established by Article 6 of the AI Act, such as the high risks systems listed in AI Act Annex III If the AI system poses a very high risk of harm to the health, safety or fundamental rights of natural persons, including by not materially influencing the outcome of decision making.	Annex III	No
Does the AI system pose a high risk of harm to the health, safety or fundamental rights of natural persons, including by not materially influencing the outcome of decision making? and /or Is the AI system is used by an Operator of Essential services (OES) and the AI system is used for the provision of an essential service as defined in Annex I of NIS 2 Directive?	Annex I of NIS 2 Directive	Yes
Does the AI system pose substantial risk of harm to the health, safety or fundamental rights of natural persons, including influencing the outcome of decision making If the AI system is used by an Operator of Essential services (OES) and the AI system is used for the provision of an important service as defined in Annex 2 of NIS2?	Annex 2 of NIS2	No
Does the AI system pose medium risk of harm to the health, safety or fundamental rights of natural persons, including influencing the outcome of decision making? Is the AI system for the provision of an important service as defined in Annex 2 of NIS2?		No
Does the AI system pose low or no risk of harm to the health, safety or fundamental rights of natural persons, including influencing the outcome of decision making?		No

Figure 29: AC Dry-run Criticality Level

The criticality assessment highlighted that, although the AC system does not fall under the high-risk categories defined in Annex III of the AI Act, it operates in a high-criticality operational context. The system is used by port authorities, which qualify as essential entities under the NIS2 Directive, and directly affects the navigation behaviour of underwater vehicles. A malfunction in the coverage algorithm may result in unsafe trajectories, including potential collisions with vessels or underwater structures. As a consequence, despite the absence of formal classification as a high-risk AI system under the AI Act, the AC component is considered high-criticality from an operational and safety perspective.

The threat assessment phase revealed a threat profile that is strongly influenced by the physical and embedded nature of the system. Threats such as dataset or validation data manipulation are considered highly unlikely due to the controlled environment in which datasets are generated and managed, including secure engineering pipelines and version-controlled data handling. Similarly, adversarial attacks such as evasion, oracle attacks, or white-box adversarial example generation are assessed as extremely rare or infeasible, given that the algorithm runs onboard the vehicle and is not exposed through external interfaces. Conversely, higher likelihood is associated with operational risks, including failures or malfunctions of the algorithm due to environmental variability, sensor noise, or hardware/software issues. Human error is also recognised as a relevant source of failure, particularly in complex deployment scenarios. In addition, risks related to component compromise and cybersecurity incidents are considered plausible, especially due to dependencies on third-party libraries and maintenance processes. Overall, the threat profile confirms that the AC system is primarily exposed to environmental uncertainty and system-level failures, rather than traditional cyber, attacks targeting externally accessible AI services.

The consequence assessment further reinforces the high criticality of the system. Safety is evaluated as having very high impact, as incorrect coverage planning may lead to collisions with ships, docks or seabed structures, potentially causing severe damage to infrastructure and vehicles, as well as risks to human operators. From a business perspective, such failures may result in costly recovery operations, delays in port activities, and disruption of scheduled inspections. From a legal perspective, liabilities may arise due to violations of safety regulations, environmental damage, or contractual obligations. Security and resilience are also assessed as very high, as compromised coverage planning could enable malicious manipulation of inspection operations or degradation of situational awareness. Explainability and interpretability are evaluated as moderate, as lack of transparency in coverage decisions may reduce operator trust and hinder troubleshooting, but does not directly result in immediate safety hazards. Privacy is considered negligible, as the system operates on environmental and infrastructure data rather than personal information. Fairness is also negligible, given the absence of human-related decision-making. Accountability and transparency are assessed as moderate, reflecting their importance in ensuring traceability of decisions and compliance with regulatory requirements. Finally, validity and reliability are identified as critical, with very high impact in case of inaccuracies or lack of robustness, as incomplete or incorrect coverage may lead to undetected hazards and inefficient mission execution.

The vulnerability analysis-based on control questions at the design stage, highlights a mixed maturity profile. On the one hand, the system demonstrates strong technical robustness. The coverage algorithm is designed using established approaches, with validation conducted in controlled environments before deployment. Datasets are generated and validated internally by

Lifecycle Stage	Affected AI Assets (these are further specified through the control questions)	Threat	Threat Definition *(from: https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms)*	Answer (use the "Default" value if you do not know the answer)	Comment UNIFI
Design	Data; Algorithm	**Area modification**	An attack in which the attacker corrupts the validation datasets.	Never	Validation sets are defined offline by the engineering team inside a secured environment; external actors have no path to tamper with areas once the dataset is version-controlled.
Testing	Algorithm; IT Systems Network and Computing	**Evasion**	A type of attack in which the attacker works on the algorithm's inputs to find small perturbations leading to large modification of its outputs (e.g. decision errors). It is as if the attacker created an optical illusion for the algorithm. Such modified inputs are often called adversarial examples.	Never	A physical underwater "illusion" is possible but would require a determined saboteur with algorithm developer access—extremely rare in our situation
Testing	Algorithm; IT Systems Network and Computing	**Use of adversarial examples crafted in white or grey box conditions**	In some cases, the attacker has access to information (algorithm, algorithm parameters, etc.) that can allow him to directly build adversarial examples.	Never	Insiders have no parameter access and nothing is learned on-board, so crafting white-box adversarial inputs has not been observed
Testing	Algorithm; IT Systems Network and Computing; Maintenance	**Oracle**	A type of attack in which the attacker explores a Algorithm by providing a series of carefully crafted inputs and observing outputs. These attacks can be previous steps to more harmful types, evasion or poisoning for example. It is as if the attacker made the Algorithm talk to then better compromise it or to obtain information about it (e.g. Algorithm extraction) or its training data (e.g. membership inferences attacks and Inversion attacks). Example: an attacker studies the set of input output pairs and uses the results to retrieve training data.	Never	The vehicle only exposes messages on an isolated Wi-Fi/buoy link; query-and-observe probing from outside the port LAN has not been reported
Production	Data; Algorithm; IT Systems Network and Computing; Maintenance	**Failure or malfunction of AI application**	This threat refers to AI application failure (e.g. denial of service due to bad input, unavailability due to a handling error). Example: the service level of the support infrastructure of the AI application hosted by a third party is too low compared to the business needs, the application is regularly unavailable. Note that this threat does not consider failure of business use cases (for example, the algorithm fails because it is not accurate enough to handle all real-life situations it is exposed to).	Once a year	Failures or malfunctions can occur due to environmental factors, hardware issues, or software bugs, making this a more frequent concern.
Production	Data; Algorithm; IT Systems Network and Computing	**Denial of service due to inconsistent data or a sponge example**	AI algorithms usually consider input data in a defined format to make their predictions. Thus, a denial of service could be caused by input data whose format is inappropriate. It may also happen that a malicious user of the Algorithm constructs an input data (a sponge example) specifically designed to increase the computation time of the Algorithm and thus potentially cause a denial of service	Never	Not feasible for this AI algorithm
Design; Testing; Production	Data; Algorithm; IT Systems Network and Computing; Maintenance	**Poisoning**	A type of attack in which the attacker altered data or Algorithm to modify the AI algorithm's behavior in a chosen direction (e.g. to sabotage its results, to insert a backdoor). It is as if the attacker conditioned the algorithm according to its motivations. Such attacks are also called causative attacks. Example: massively indicating to an image recognition algorithm that images of dogs are indeed cats to lead it to interpret it this way.	Never	The AI algorithm is runned onboard the vehicle with almost impossible poisining issues
Design; Testing; Production	Data; Algorithm; IT Systems Network and Computing; Maintenance	**Algorithm or data disclosure**	This threat refers to the possibility of leakage of all or partial information about the Algorithm.12 Example: the outputs of a AI algorithm are so verbose that they give information about its configuration (or leakage of sensitive data)	Once every 10 years	Only coverage % and depth grids are transmitted; historical sector reviews show very rare accidental disclosure events
Design; Testing; Production	Data; Algorithm; IT Systems Network and Computing	**Data disclosure**	This threat refers to a leak of data manipulated by AI algorithms. This data leakage can be explained by an inadequate access control, a handling error of the project team or simply because sometimes the entity that owns the Algorithm and the entity that owns the data are distinct. To train the Algorithm, it is often necessary for the data to be accessed by the Algorithm provider. This involve sharing the data and thus share sensitive data with a third party.	Once every 10 years	Inspection recordings may be classified as "port-security sensitive" and reside on secure server; history shows only very occasional accidental exposure via misplaced removable media.
Design; Testing; Production	Algorithm; IT Systems Network and Computing	**Algorithm disclosure**	This threat refers to a leak of the internals (i.e. parameter values) of the AI Algorithm. This Algorithm leakage could occur because of human error or contraction with a third party with a too low security level.	Once every 5 years	The AI algorithm parameters are stored in the vehicle but cloud back-ups may exist with very rare leakage possibility
Design; Testing; Production	Algorithm; IT Systems Network and Computing; Maintenance	**Compromise of AI application components**	This threat refers to the compromise of a component or developing tool of the AI application. Example: compromise of one of the open-source libraries used by the developers to implement the AI algorithm.	Once a year	External libraries may require update or patching; typical disclosure cadence for similar robotics applications is every 12-18 months.
Design; Testing; Production	Algorithm; IT Systems Network and Computing; Maintenance	**Human error**	The different stakeholders of the Algorithm can make mistakes that result in a failure or malfunction of AI application. For example, due to lack of documentation, they may use the application in use-cases not initially foreseen.	Once a year	End-user errors can occur, especially in complex systems like underwater drones, leading to occasional failures or malfunctions.
Design; Testing; Production	IT Systems Network and Computing; Maintenance	**Cybersecurity incident not reported to incident response teams**	This threat refers to the possibility that a project team may not report security incidents to dedicated teams while a policy of mandatory incident reporting has been defined.	Once every 2 years	Underreporting of cybersecurity incidents is a known issue, and while efforts are made to report them, some incidents may still go unreported.

Figure 30: AC Dry-run Frequency of Threats

expert engineers, and data quality is ensured through multi, level verification processes.

Performance indicators for proper functioning are defined, and simulation-based validation is

Trustworthiness characteristic	Question	Answer	Please provide in detail potential business and legal consequences UNIFI
Safety	What is the business consequence level if the **safe** operation of the AI system is compromised?	Very High	Business: Collision with ships, docks or seabed structures could sink or disable the AUV and/or damage port assets, forcing emergency port closures, dredging delays and re-survey missions that raise direct costs (vessel hire, diver intervention) and cause cascading schedule disruptions for port clients. Legal: the AUV system integrator may face liability claims for property damage and environmental harm, breach of port-safety regulations and higher insurance premiums; a serious incident could trigger suspension of operating permits by the Port Authorities and fines under national maritime-safety statutes.
Security and resilience	What is the business consequence level if the **security or the resilience** of the AI system to adverse events is compromised?	Very High	Business: A cyber-attack or sensor spoofing that alters the planned track or falsifies bathymetric data could oblige UNIFI to repeat surveys, eroding client confidence and delaying downstream dredging or construction projects. Legal: Ports can invoke contractual penalty clauses, while data-integrity breaches fall under EU NIS2 obligations with potential administrative fines.
Explainability and interpretability	What is the business consequence level if the AI system is not **interpretable**?	Moderate	Business: If operators cannot see why the CPP algorithm chose a path or altered it in real time, they may override or abandon the tool, cutting AC service revenues. Legal: under the AI Act, high-risk AI used in critical infrastructure must provide "meaningful information" on logic and decisions; non-compliance exposes UNIFI to market-surveillance actions.
	What is the business consequence level if the AI system is not **explainable**?	Moderate	As before
Privacy	What is the business consequence level caused by a **privacy** compromise of the AI system?	Negligible	The AC system does not process personal data; residual privacy risk is limited to incidental capture of divers or vessel identifiers in multibeam records.
Fairness with harmful bias managed	What is the business consequence level if the AI system is not **fair**?	Negligible	In this context, fairness is negligible
Accountability and Tranparency	What is the business consequence level if the AI system is not **transparent**?	Moderate	Business: Missing audit trails or tamper-proof logs make it hard to defend AC decisions after an incident, exposing UNIFI to protracted disputes, reputational loss and higher professional-indemnity costs. Legal: Maritime authorities can withhold operating licences until traceability is demonstrated
Validity and reliability	What is the business consequence level if the AI system is not **accurate**?	Very High	Business: Inaccurate bathymetric maps create false alarms, disrupt vessel scheduling and may let real seabed hazards go unnoticed, forcing costly repeat missions and downtime. Legal: systematic errors expose UNIFI to damage claims
	What is the business consequence level if the AI system is not **robust**?	Very High	Business: Sensor faults, calibration drift or CPP failures in dynamic port conditions lead to incomplete coverage and wasted AUV time; frequent mission aborts undermine maintenance budgets and client confidence. Legal: Persistent robustness issues can be deemed an "unsafe product"

Figure 31: AC Dry-run Business Consequence Level

used to reduce risks before real-world deployment. On the other hand, several gaps are identified in terms of security and governance controls. In particular, mechanisms related to access control, identity management, formal integration of security into the development lifecycle, and systematic vulnerability assessment of components are not fully implemented. Similarly,

Lifecycle Stage	Threat(s)	Vulnerability (from Golden Table)	Control	Relevant AI Asset for the control question (Replace with the AI Assets of your SUT)	Template Control Questions	Answer	Comment (please add here your comments about the questions)
Design	**Dataset modification**	Use of unreliable sources to **define validation dataset**	Ensure reliable sources are used	Dval1	*In the context of Dval, have you ensured that reliable sources were used to define the areas inside the dataset?*	Yes	Areas defined in-house by expert underwater robotics engineers with a minimum 2-person quality check
Design; Testing; Production	**Poisoning**	Algorithm vulnerable to data poisioning adversarial attacks	Choose and define a more resilient algorithm design	Alg1	*In the context of Alg1, have you chosen and defined a resilient algorithm design?*	Yes	planner; field trials show graceful recovery under sensor noise or dynamic obstacles
			Assess the exposure level of the algorithm used	Alg1	*In the context of Alg1, have you assessed the exposure level of the algorithm to poisoning adversarial attacks?*	No	
	Poisoning	Lack of data for increasing robustness to data poisoning attacks	Enlarge the training dataset (by adding more diverse data)	Alg1	*In the context of M1, have you enlarged the validation dataset by adding diverse data to increase its robustness to poisoning attacks?*	Yes	Areas from simulated experience rather on-field trials are added to increase algorithm robustness
	Poisoning, algorithm or data disclosure, Compromise of AI application	Poor access rights management	Apply a RBAC algorithm, respecting the least privileged principle	Alg1	*In the context of Alg1, have you applied an RBAC algorithm, respecting the least privileged principle?*	No	
				Dval1	*In the context of Tr1, have you applied an RBAC algorithm, respecting the least privileged principle?*	No	
	Poisoning	Poor data management	Ensure AI applications comply with data security requirements	Alg1	*In the context of Alg1, have you ensured that the algorithm complies with data security requirements?*	No	
	Compromise of AI application components, Poisoning	Undefined indicators of proper functioning, making complex compromise identification	Define and monitor indicators for proper functioning of the algorithm	Alg1	*In the context of Alg1, have there been defined indicators of proper functioning of the algorithm?*	Yes	Standard CPP metrics
	Data disclosure, Compromise of AI application components,	Lack of consideration of attacks to which AI applications could be exposed to	Integrate AI specificities to awareness strategy and ensure all AI stakeholders are receiving it	Alg1	*In the context of Alg1, have you integrated AI specifities to awareness strategies and ensured all AI stakeholders are receiving it?*	No	
	Human error, Poisoning	Use of uncontrolled data	Control all data used by the AI algorithm	Alg1	*In the context of Alg1, have you controlled all data used by the algorithm?*	Yes	Data controlled by expert underwater robotics engineers with a minimum 2-person quality check
	Poisoning	Use of unsafe data or algorithms (e.g. with transfer learning)	Ensure reliable sources are used	Dval1	*Is Dval reliable?*	Yes	The validation areas are tested with Alg1 in a simulated environment to prevent errors
				Alg1	*Is the base algorithm used for Alg1 reliable?*	Yes	OMPL's RRT* implementation is open-source, unit-tested and sea-trial-validated
	Poisoning	No detection of adversarial samples in the training dataset	Use methods to clean the training dataset from adversarial samples	Dval1	*Does Dtr include adversarial samples?*	No	
	Poisoning, algorithm or data disclosure, Compromise of AI application components	Weak access protection mechanisms for AI algorithm components	Ensure AI applications comply with identity management, authentication, and access control policies	Alg1	*In the context of Alg1, have you ensured that the algorithm complies with with identity management, authentication, and access control policies ?*	No	
	algorithm or data disclosure, Compromise of AI application components	Existence of unidentified disclosure scenarios	Conduct a risk analysis of the AI application	Alg1	*In the context of Alg1, have you conducted risk analysis to identify all disclosure scenarios?*	No	
	algorithm or data disclosure,	Lack of security process to maintain a good security level of the components of	Implement processes to maintain security levels of AI components over time	Alg1	*Have you implemented processes to maintain security levels of Alg1 components over time?*	No	
	Data disclosure	The algorithm can allow private information to be retrieved	Ensure that algorithms respect differential privacy.	Alg1	*Does the Alg1 algorithm respect differential privacy?*	No	
			Reduce the information given by the algorithm	Alg1	*Have you designed the Alg1 algorithm in such a way that will provide only the necessary information and will not disclose any private information?*	No	
	Data disclosure,	Disclosure of sensitive data for AI algorithm training	Use federated learning to minimise the risk of data breaches	Alg1	*In the context of Alg1, have you used ferated learning to minimise the risk of data breaches?*	No	
	Data disclosure, algorithm disclosure, Compromise of AI application components	Too much information available on the algorithm architecture	Limit the available information about the algorithm architecture	Alg1	*In the context of Alg1, have you limited the available information about the algorithm architecture?*	Yes	The algorithm is known and available from our scientific literature, but the parameter tuning and the specific adaptation is kept internal
	Data disclosure, algorithm disclosure, Compromise of AI application components	Too much information about the algorithm given in its outputs	Reduce the output information given by the algorithm	Alg1	*In the context of Alg1, have the output information given by the algorith*	Yes	Operator console displays aggregate coverage %, not full state tree, thus minimising leakage
	Compromise of AI application components	Existence of several vulnerabilities because the AI application was not included into process for integrating	Ensure AI projects follow the global process for integrating security into projects	Alg1	*In the context of Alg1, have it been ensured that the algorithm has followed the global process for integrating security?*	No	
	Compromise of AI application components	Use of vulnerable components (among the whole supply chain)	Check the vulnerabilities of the components used so that they have an appropriate security level	Alg1	*In the context of Alg1, have you checked the vulnerabilities of the components used so that they have an appropriate security level?*	No	
	Compromise of AI application components	Existence of several vulnerabilities because AI specificities are not integrated to existing policies	Integrate AI specificities to existing security policies	Alg1	*In the context of Alg1, have there been integrated algorithm specifities to existing security policies?*	No	
	Compromise of AI application components	Existence of several vulnerabilities because AI application do not comply with security policies	Ensure AI applications comply with security policies	Alg1	*In the context of Alg1, have you ensured that the algorithm complies with the relevant security policies?*	No	
			Include AI applications into asset management processes	Alg1	*Have you included the Alg1 algorithm into the asset management processes?*	No	
	Compromise of AI application components	Contract with a low security third party	Ensure AI applications comply with third parties' security requirements	Alg1	*Have you ensured that the Alg1 algorithm complies with third parties' security requirements?*	No	

Figure 32: AC Dry-run Control Questions Design

advanced measures such as federated learning, differential privacy, or structured processes for

managing disclosure risks are currently absent. The algorithm architecture is partially exposed through scientific publications, although implementation details and parameter tuning remain internal. Output information is limited to aggregated indicators (e.g., coverage percentage) to reduce potential leakage, which represents a positive mitigation measure.

3.2.2 Plan development

The planning activities for the validation and deployment of FAITH tools within the LSP4 underwater robotics pilot were structured and consolidated during the Oslo plenary meeting, where APRA, UNIFI and VPF jointly defined a detailed implementation roadmap for the period covering M23 to M36, with further continuation of some contextual modelling and replication-related activities beyond this window. The purpose of this planning effort was to translate the methodological principles of the FAITH AI_TAF into a concrete and operational validation plan, aligned with the maturity level of the pilot and with the expected progression from development to deployment in real port environments. The plan was organised around two main macro-phases. The first corresponds to a development and preparation phase, beginning in M23, in which the partners define the validation strategy, receive the necessary training on FAITH tools, identify the relevant assets and users, and prepare the conditions for tool adoption. The second corresponds to a deployment and iterative validation phase, beginning progressively from M26-M27, in which the tools are applied to the actual SUT, first in the initial pilot at the Port of Ravenna and later in the replication pilot at the Port of Valencia.

The first planning actions were scheduled in M25, when the consortium formalised the pilot-specific implementation plan. As of M29, these initial actions have been fully completed: both Define plan and Agree plan are marked at 100% progress, confirming that the roadmap was formally established and shared among the pilot partners. A major component of the plan concerns training and support, which was distributed over several months to ensure progressive familiarisation with the different tools. The training on trustSENSE was planned already in M23, reflecting the importance of early preparation on team maturity and human-factor assessment, and this activity is now completed (100%). Training on the AI Model Passport and AI Data Passport tools was scheduled between M26 and M27, allowing the partners to prepare model and data documentation before or during the first deployment-related activities; both training actions are now completed (100%). Training on trustGUARD was also scheduled between M26 and M27, in order to support the first rounds of model and data asset assessment, and this too is completed (100%). In contrast, the training on the SSM and the introduction to the FAITH STM Platform registration and workflows were planned later, between M34 and M35, coherently with the fact that these tools are used at a more advanced stage, when the pilot has already generated sufficient evidence and maturity for system-level and in-context assessment. At the current stage, both activities remain pending (0%).

Detailed Implementation Plan LSP4 - Underwater Robotics		Phase:	Development	PIs delete 1:	Full Deployment				*NOTE: You will need to answer trustGUARD questions as if the SUT is deployed in reality (not in the sandbox)									
APRA																		
	Project Start:		M23															
	Display Month:		1		M23	M24	M25	M26	M27	M28	M29	M30	M31	M32	M33	M34	M35	M36
TASK	ASSIGNED TO	PROGRESS	START	END														
Planning																		
Define plan	APRA / UNIFI / VPF	100%	M25	M25														
Agree plan	APRA / UNIFI / VPF	100%	M25	M25														
Training and support																		
Attend training for AI Model Passport (pythonic library)	APRA / UNIFI / VPF	100%	M26	M27														
Attend training for AI Data Passport	APRA / UNIFI / VPF	100%	M26	M27														
Attend training for trustSENSE	APRA / UNIFI / VPF	100%	M23	M23														
Attend training for trustGUARD	APRA / UNIFI / VPF	100%	M26	M27														
Attend training for SSM	APRA / UNIFI / VPF	0%	M34	M35														
Attend introduction to FAITH STM Platform registration and workflows	APRA / UNIFI / VPF	0%	M34	M35														
Deploy SUT in Port or Ravenna Port	APRA / UNIFI	100%	M27	M28														
Use trustSENSE for team maturity assessment (early deployment period)	Risk Assessor + End Users	100%	M26	M26														
Provide formal feedback on Usability and Usefulness of trustSENSE @ deployment phase	Risk Assessor	100%	M26	M26														
Implement trustSENSE mitigations for end-users @deployment phase	APRA/VPF	10%	M26	M36														
Create asset entries in AI Model Hub	Risk Assessor	100%	M26	M28														
Provide formal standardised feedback on Usability and Usefulness of AI Model Hub	Risk Assessor	100%	M28	M28														
Iterate AI Model Hub v2 based on LSP feedback	FORTH / CNR	0%	M26	M32														
Use trustGUARD v1 for in-isolation AI Model Asset assessment @ deployment phase	Risk Assessor	100%	M27	M28														
Use trustGUARD v1 for in-isolation (Input) Data Asset assessment @ deployment phase	Risk Assessor	100%	M27	M28														
Provide formal standardised feedback on Usability and Usefulness of trustGUARD v1 during deployment phase	Risk Assessor	100%	M29	M29														
Apply 1 or more Controls to AI Model Asset or Data Asset during deployment phase based on trustGUARD recommendations	UNIFI	10%	M29	M36														

Figure 33: Example of the developed plan

The planning phase also defined the transition from preparation to operational deployment. The deployment of the SUT in the Port of Ravenna was scheduled across M27 and M28, with APRA and UNIFI sharing responsibility. As of M29, this deployment activity is marked at 100% progress, which means that the first deployment cycle in Ravenna has been completed and has effectively constituted the beginning of the early operational validation stage, enabling the first application of FAITH tools on the pilot in a real port environment. Prior to and around this transition, several assessment activities were scheduled in M26, especially those related to trustSENSE. More specifically, the use of trustSENSE for team maturity assessment of end users during the early deployment period was planned in M26, and this activity is now completed (100%), together with the collection of formal feedback on the usability and usefulness of trustSENSE in the same month (100%). The plan also foresaw the implementation of trustSENSE-based mitigations for end users from M26 to M36, meaning that the outputs of the tool were not intended to remain descriptive only, but to feed continuous improvement actions over a long period. At the present stage, this mitigation activity has started but remains at an early level (10%).

In parallel, the plan integrated documentation and traceability activities through the AI Model Hub. The creation of asset entries in the AI Model Hub was scheduled from M26 to M28, and this activity has now been completed (100%). The collection of formal feedback on the usability and usefulness of the AI Model Hub, planned in M28, is also completed (100%). The plan then foresaw

a longer-term iterative refinement of the platform, through the activity Iterate AI Model Hub v2

Update and re-run in-isolation AI Model Asset and/or Data Asset assessment including new Control(s) @ deployment phase using trustGUARD v1	Risk Assessor	0%	M32	M36														
Re-train SUT with extended image datasets from Ravenna Port (design phase)	UNIFI	100%	M27	M28														
Use trustSENSE for team maturity assessment (@ design phase -> AI Technical Team)	Risk Assessor +UNIFI	100%	M26	M26														
Provide formal feedback on Usability and Usefulness of trustSENSE @ design phase	Risk Assessor	100%	M26	M26														
Implement trustSENSE mitigations for AI Technical Team @ design phase	UNIFI	10%	M27	M36														
Use trustGUARD v1 for in-isolation AI Model Asset assessment @ design phase	Risk Assessor	10%	M29	M29														
Use trustGUARD v1 for in-isolation (Input) Data Asset assessment @ design phase	Risk Assessor	0%	M29	M29														
Provide formal standardised feedback on Usability and Usefulness of trustGUARD v1 during design phase	Risk Assessor	0%	M29	M29														
Apply 1 or more Controls to AI Model Asset or Data Asset during design phase based on trustGUARD recommendations	APRA / UNIFI	0%	M29	M36														
Update and re-run in-isolation AI Model Asset and/or Data Asset assessment including new Control(s) @ design phase using trustGUARD v1	Risk Assessor	0%	M32	M33														
Use FAITH STM Platform	Risk Assessor	0%	M35	M35														
Provide formal standardised feedback on Usability and Usefulness of FAITH STM Platform	Risk Assessor	0%	M36	M36														
Iterate FAITH STM Platform	??? + UOS	0%	M34	M36														
Create SSM system model	APRA / UNIFI / UOS	0%	M33	M33														
Use SSM for in-context SUT assessment	Risk Assessor	0%	M33	M35														
Provide formal standardised feedback on Usability and Usefulness of SSM	Risk Assessor	0%	M35	M35														
On-going interation of SSM v2 based on LSP feedback (inc. on-going Knowledgebase updates)	UOS	0%	M36	M45														
Deploy SUT in Port of Valencia (Replication Pilot)	VPF / UNIFI	0%	M33	M45														

Figure 34: Example of developed plan

based on LSP feedback, running from M26 to M32, under the responsibility of FORTH/CNR. This demonstrates that the Oslo planning did not treat model and data documentation as a one-off compliance exercise, but rather as an evolving component of the pilot validation process. However, at the current checkpoint this longer-term refinement activity is still marked at 0%.

A particularly relevant block of activities in the implementation plan concerns the use of trustGUARD. For the deployment phase, the plan scheduled the use of trustGUARD v1 for in-isolation AI Model Asset assessment and trustGUARD v1 for in-isolation Input/Data Asset assessment between M27 and M28, under the responsibility of the risk assessor. As of M29, both of these assessment activities are completed (100%). The collection of formal standardised feedback on the usability and usefulness of trustGUARD v1 during deployment, scheduled for M29, is also completed (100%). Based on the recommendations generated by trustGUARD, the plan foresaw the application of one or more additional controls to AI model or data assets from M29 to M36, under the responsibility of UNIFI. This activity has now started, but remains at an initial stage (10%). It is complemented by an explicit re-assessment step, namely the update and re-run of the in-isolation AI model and/or data asset assessment including new controls, scheduled from M32 to M36. At the current stage this re-assessment has not yet started (0%). This sequence shows that the implementation plan adopts an iterative loop: first assessment, then feedback, then corrective action, then re-assessment.

The plan developed in Oslo also preserved a clear distinction between deployment-phase assessment and design-phase assessment, which is important in the FAITH methodology. For the design phase, trustSENSE was again scheduled in M26, this time for the AI Technical Team maturity assessment, followed by usability and usefulness feedback in the same month. At the current stage, both of these activities are completed (100%). The implementation of trustSENSE-based mitigations for the AI Technical Team was planned from M27 to M36, mirroring the long-term nature of organisational improvement, and is currently underway at an early level (10%). Similarly, trustGUARD v1 for in-isolation AI Model Asset assessment at design phase and trustGUARD v1 for in-isolation Input/Data Asset assessment at design phase were planned in M29. Here the current status is more partial: the AI Model Asset assessment is marked at 10%, while the Input/Data Asset assessment remains at 0%. The associated formal standardised feedback on the usability and usefulness of trustGUARD v1 during design phase, also scheduled for M29, remains at 0%, as does the subsequent application of controls from M29 to M36 and the update and re-run of these design-phase in-isolation assessments with new controls, planned for M32-M33. This distinction between design and deployment assessments remains methodologically significant because it allows the consortium to compare how trustworthiness evolves when moving from the engineering environment to real operations, even if part of this design-phase track is still in progress or pending at the current checkpoint.

The implementation plan also includes a specific technical improvement action for the perception pipeline, namely the re-training of the SUT with extended image datasets from Ravenna Port, scheduled between M27 and M28. This activity is now completed (100%). Even though this action is more directly linked to ATR, its presence in the roadmap shows that the planning process integrated assessment and technical refinement rather than treating them as separate streams. This is fully coherent with the FAITH philosophy, in which assessment results are expected to drive concrete system improvements.

The later months of the plan, from M33 onwards, mark the transition from asset-level trustworthiness assessment to system-level and in-context analysis. The creation of the SSM system model is planned in M33, involving APRA, UNIFI and UOS, while the use of SSM for in-context SUT assessment is planned from M33 to M35. Feedback on the usability and usefulness of SSM is scheduled in M35, and the on-going iteration of SSM v2 based on LSP feedback, including on-going knowledge-base updates, is planned from M36 to M45, under the responsibility of UOS. In addition, the dedicated use of the FAITH STM Platform is scheduled in M35, followed by feedback collection in M36 and platform iteration from M34 to M36. At the present stage, all these activities remain pending (0%). These activities demonstrate that the Oslo planning was not limited to early dry-run tools, but extended to the more mature FAITH components supporting contextual and system-wide modelling. Finally, the plan explicitly

includes the replication pilot at the Port of Valencia, scheduled from M33 to M45, with VPF and UNIFI responsible for deployment; this activity also remains pending (0%).

From a methodological standpoint, one of the strongest features of the plan defined during the Oslo plenary is its overlapping and iterative temporal structure. Activities are not organised as a rigid linear chain. Rather, training, deployment, assessment, feedback collection, mitigation, re-assessment and contextual modelling partially overlap across M23-M36, allowing lessons learned in one phase to immediately influence subsequent actions. At the current M29 checkpoint, this structure is already visible in practice: the pilot has completed the core planning, tool training, initial Ravenna deployment, first trustSENSE assessments and feedback, AI Model Hub registration and feedback, trustGUARD deployment-phase assessments, and re-training of the SUT, while mitigation actions have already started and the later re-assessment and system-level modelling steps remain ahead in the roadmap.

Although the validation and implementation plan described in this section was developed primarily with the ATR branch in mind, it also reflects, from a procedural and organisational point of view, the overall approach that will be followed for the AC branch. In particular, the same logic of progressive preparation, tool training, asset identification, in-isolation assessment, feedback collection, mitigation implementation, and iterative re-assessment applies to both AI systems, even where the specific technical assets, datasets and validation evidence differ between ATR and AC.

3.2.3 Tool Training

TrustSENSE

The training activities related to trustSENSE were initiated in M23 and were formally supported through the organisation of the 2nd FAITH Workshop, dedicated specifically to the trustworthiness assessment of AI participants teams. The session was held online on 10 November 2025, from 18:00 to 20:30 (GMT+3), and was organised by Trustilio as a dedicated FAITH training and validation event focused on familiarising project partners with the trustSENSE tool, its assessment logic, and its intended operational use within the project.

From a methodological perspective, the trustSENSE training had a dual role. First, it served as a capacity, building activity, ensuring that pilot partners and relevant stakeholders developed a shared understanding of the tool before its use in the validation activities. Second, it functioned as an early validation exercise for the tool itself, by checking whether its structure, scoring outputs and mitigation logic were understandable, interpretable and practically relevant across different user roles. The workshop objectives explicitly included ensuring that participants fully understood the trustSENSE tool, its domains, scoring logic, and mitigation linkages; testing whether the produced scores and proposed mitigations were interpretable and role-relevant; verifying

whether the weighting of the scoring dimensions matched expert judgment across different responder profiles; and confirming the overall usability, clarity and practical adoption potential of the tool.

The training content was organised through a structured agenda that reflects a progressive learning approach. After an initial welcome and introduction to the workshop objectives, the session moved to a dedicated trustSENSE presentation, in which the conceptual foundations and operational logic of the tool were introduced. This was followed by a case-based scoring exercise by responder type, which is particularly relevant for FAITH because trustSENSE is designed to assess the trustworthiness, related maturity of human teams involved in the AI lifecycle, rather than only the technical properties of AI assets. The workshop then addressed scoring weighting and sensitivity discussion, allowing participants to reflect on how different factors contribute to final scoring outcomes and whether these outputs align with expert expectations. The final technical block focused on usability and human-centric fit, before moving to a broader discussion and wrap, up. This sequence indicates that the training did not simply explain how the tool works, but also actively engaged participants in evaluating its interpretability and applicability in realistic organisational contexts.

Within the context of the LSP4 underwater robotics pilot, this training was particularly relevant because the trustworthiness of the AI systems under assessment depends not only on models, datasets and algorithms, but also on the maturity, awareness and preparedness of the human teams interacting with them. In this respect, trustSENSE provides a structured way to assess the readiness of AI participants teams, including technical teams and operational users, and to identify mitigations aimed at improving organisational and human-related trustworthiness conditions. The training therefore supported the subsequent use of trustSENSE in both design, phase and deployment, phase activities, by preparing partners to correctly interpret assessment domains, scoring profiles and mitigation recommendations before applying the tool to the specific pilot setting. This is fully coherent with the broader FAITH approach, in which technical and human dimensions of trustworthiness are jointly considered throughout the assessment process.

Data & AI Model Passport in the FAITH AI Model Hub

The training on the Data Passport and AI Model Passport within the FAITH AI Model Hub was carried out in M26–M27 and was especially important for LSP4 because it taught the partners not only what these tools are for, but how to use them concretely in the underwater robotics pilot. The main goal of the training was to enable the consortium to document datasets, models and development pipelines in a structured way, so that the assets used in ATR and, progressively, in

the broader pilot workflow could be traced, understood and assessed consistently throughout the project.

For the Data Passport, the training, delivered on a virtual meeting by FORTH, showed the partners how to create a structured digital identity for a dataset by completing the main documentation fields related to dataset description, authorship, motivation, intended and non-intended use, lifecycle role, distribution and security measures. In practice, this meant learning how to record where a dataset comes from, who owns it, what it can be used for, how it was produced and annotated, under which access conditions it is shared, and which legal, organisational and technical protections apply. For LSP4, this was particularly relevant because the underwater robotics pilot relies on pilot-specific datasets acquired in real port environments, and these data are directly linked to safety-relevant AI functions. The training therefore helped ensure that datasets were not treated as informal technical resources, but as governed assets with explicit provenance, use boundaries and traceability.

For the AI Model Passport, the training had a more practical and pipeline-oriented character. The partners were taught how to use the passport library inside the FAITH environment, how to initialise the passport workflow, generate the required configuration files, connect the model to the corresponding Data Passport through the dataset UID, and execute the passport process across the relevant pipeline stages. In other words, the training explained how to make the model-development workflow itself traceable, from preprocessing and training to evaluation and production. This was important for LSP4 because the pilot does not only require documentation of final models, but also the possibility to reconstruct how those models were generated, which data supported them, and under which pipeline conditions they were produced.

Another important aspect of the training is that it clarified the link between data and model documentation. The partners were taught that the Data Passport and AI Model Passport should not be used as separate isolated records, but as connected elements of the same trustworthiness chain. This is particularly valuable for the underwater robotics pilot, where model behaviour can only be properly interpreted if the supporting dataset, the intended use, the provenance and the experimental setup are all explicitly documented. The training therefore helped prepare LSP4 not only for technical onboarding to the platform, but also for a more rigorous governance of AI assets.

TrustGuard

The training activities related to TrustGuard were carried out in M26–M27, in line with the implementation plan agreed during the Oslo plenary, and were specifically organised to prepare the LSP4 risk assessors and technical partners for the practical use of the tool within the underwater robotics pilot. The TrustGuard training session with LSP4 was scheduled on 18

February 2026, from 10:00 to 11:00, in virtual format via Microsoft Teams, and that the purpose of the session was to train the risk assessors of LSP4 to use TrustGuard. The session was not limited to a general introduction: it was explicitly organised to teach the LSP4 partners how to perform an actual FAITH trustworthiness assessment on their pilot assets, and therefore represented the practical bridge between the methodological planning of the assessment and its real execution on ATR and AC-related assets.

The training was delivered by ATC, and was accompanied by supporting documentation, including the presentation slides, a recording of the session, and an evaluation form for follow-up feedback. The session clarified how to start an assessment, identify and select the relevant AI assets, assign the lifecycle stage, complete the criticality questions, and then proceed to the actual assessment tasks for each asset. In practice, the training taught the LSP4 partners how to move from an abstract system description to an asset-level trustworthiness assessment that could later be used in the pilot for concrete models and datasets. This was essential because, in LSP4, the correct use of TrustGuard depended on being able to assess not just “the system” in general, but specific pilot assets such as ATR models, ATR datasets and the AC asset in a structured and comparable way. The training was also very important because it clarified the internal logic of the assessment questions. Partners were shown how to reason about the frequency of threats, how to evaluate the consequences of compromised trustworthiness properties, and how to answer the vulnerability questions related to the implementation status of controls and safeguards. In this sense, the training did not only show where to click in the tool, but taught the users how to think through the assessment in a way that was aligned with the FAITH methodology.

A further important contribution was the explanation of how TrustGuard derives risk. The training clarified that risk is not based on a single answer, but on the combined effect of threat level, consequence level, vulnerability level, and maturity level, applied to specific assets and trustworthiness dimensions. For LSP4, this was essential because it helped the partners understand how weaknesses in datasets, models, software components or engineering practices could propagate into concrete trustworthiness risks affecting, for example, Validity and Reliability, Safety, Security and Resilience, or Accountability and Transparency. In other words, the training enabled LSP4 to interpret TrustGuard outputs not as isolated scores, but as the result of a structured and meaningful assessment chain.

3.3 FAITH Tool Validation Activities

3.3.1 TrustSense

Within the FAITH tool validation activities, UNIFI used trustSENSE to assess the maturity and trustworthiness readiness of its AI Technical Team, i.e. the team responsible for the development, validation and technical governance of the underwater robotics AI systems. In this case,

trustSENSE was applied as a structured self, assessment tool for the human team asset, rather than for the AI models or datasets. The assessment was completed through the trustSENSE questionnaire and produced both an overall trust level and a set of category, specific mitigation recommendations.

The outcome of the assessment classified the UNIFI AI Technical Team at a High trust level, with a calculated score of 3.80. The result page explicitly states that the team largely adheres to trustworthiness standards, with minor areas for improvement, and recommends targeted refreshers on trustworthiness dimensions, improved access to updated tools and resources, and stronger knowledge exchange with high-performing teams. In practical terms, this means that the technical team already shows a solid maturity profile for developing and managing the ATR and AC systems, but that some organisational and compliance, related aspects still require consolidation.

The scoring profile shows that the strongest dimensions of the team are Proactivity and Threat Awareness (4.00), Responsibility and Ethics (4.00), Collaboration (4.00), Technical Proficiency (4.00), Problem Solving (4.00), and Motivation and Commitment (4.00), all classified at High level. This result is coherent with the technical role of UNIFI within the pilot. It indicates that the team is perceived as proactive in identifying AI, related threats, committed to trustworthy AI principles, technically capable of handling demanding AI development tasks, and effective in collaborative problem solving. It also suggests that the team has a strong internal motivation to improve trustworthiness-oriented practices and to engage constructively with other actors involved in the pilot.

At the same time, the assessment highlighted a number of dimensions with Moderate scores, namely Innovation and Adaptability (3.50), Resilience (3.00), Integrity (3.00), Resource Accessibility (3.00), Policy Adherence (3.00), and Privacy and Compliance (3.00). These scores do not indicate critical weaknesses, but rather areas where the team's current maturity is less consolidated than in the strongest categories. In particular, the moderate values on policy adherence, privacy and compliance, and resilience suggest that the team is technically strong but can still improve in the formalisation of compliance-oriented routines, continuity and recovery practices, and systematic integration of trustworthiness procedures into day, to, day engineering work.

The detailed questionnaire responses are consistent with this overall result. Most answers were in the "4 – Agree" range, confirming a generally positive maturity profile. Strong responses were recorded on issues such as role clarity, incident mitigation, routine implementation of mitigation actions, collaboration, ethical commitment, use of advanced technical tools, understanding of multidimensional AI trustworthiness requirements, knowledge of intended system use, openness

to external feedback, routine identification of AI threats, adherence to trustworthiness practices under pressure, and data-quality management. By contrast, more neutral responses appeared in questions related to policy compliance, resilience to technological failures, legal and ethical consistency in operations, recovery within strict continuity timelines, privacy and legal compliance targets, and the formalisation of post, error process revision. This pattern explains why the overall trust level is high but not yet fully mature in every dimension.

An important feature of trustSENSE is that it does not only produce a score, but also generates category, specific mitigation actions. For the UNIFI AI Technical Team, the tool recommended maintaining the strong threat, awareness culture through monthly threat, anticipation workshops and clearer role assignment for risk monitoring. In the area of responsibility and ethics, it suggested embedding ethical role expectations into performance reviews and rotating responsibility for monthly reviews. For innovation and adaptability, the proposed action was to launch idea, generation sprints for trustworthiness improvements and assign team members to review emerging practices. For resilience, the tool recommended introducing resilience checkpoints into project plans, together with explicit roles for recovery testing and incident documentation. For collaboration, it proposed regular knowledge, exchange routines with external teams or organisations. For integrity, it recommended scenario-based discussions on integrity challenges and a visible statement of the team's ethical commitments. For technical proficiency, it suggested delegating mentoring roles for AI validation and error detection, with clearer quality, check targets. For problem solving, it recommended post, mortem reviews and the creation of a registry of solved cases. For resource accessibility, it proposed reassessing tool usage, encouraging participation in external briefings, and assigning liaison roles for resource coordination. For policy adherence, the suggested action was to organise refresher briefings on updated policies and reinforce learning through mini, quizzes or internal challenges. For motivation and commitment, it recommended rotating motivational leadership roles and introducing recognition mechanisms for trustworthiness efforts. Finally, for privacy and compliance, it suggested regular briefings with compliance officers and assigning explicit privacy, review roles during project lifecycle checkpoints.

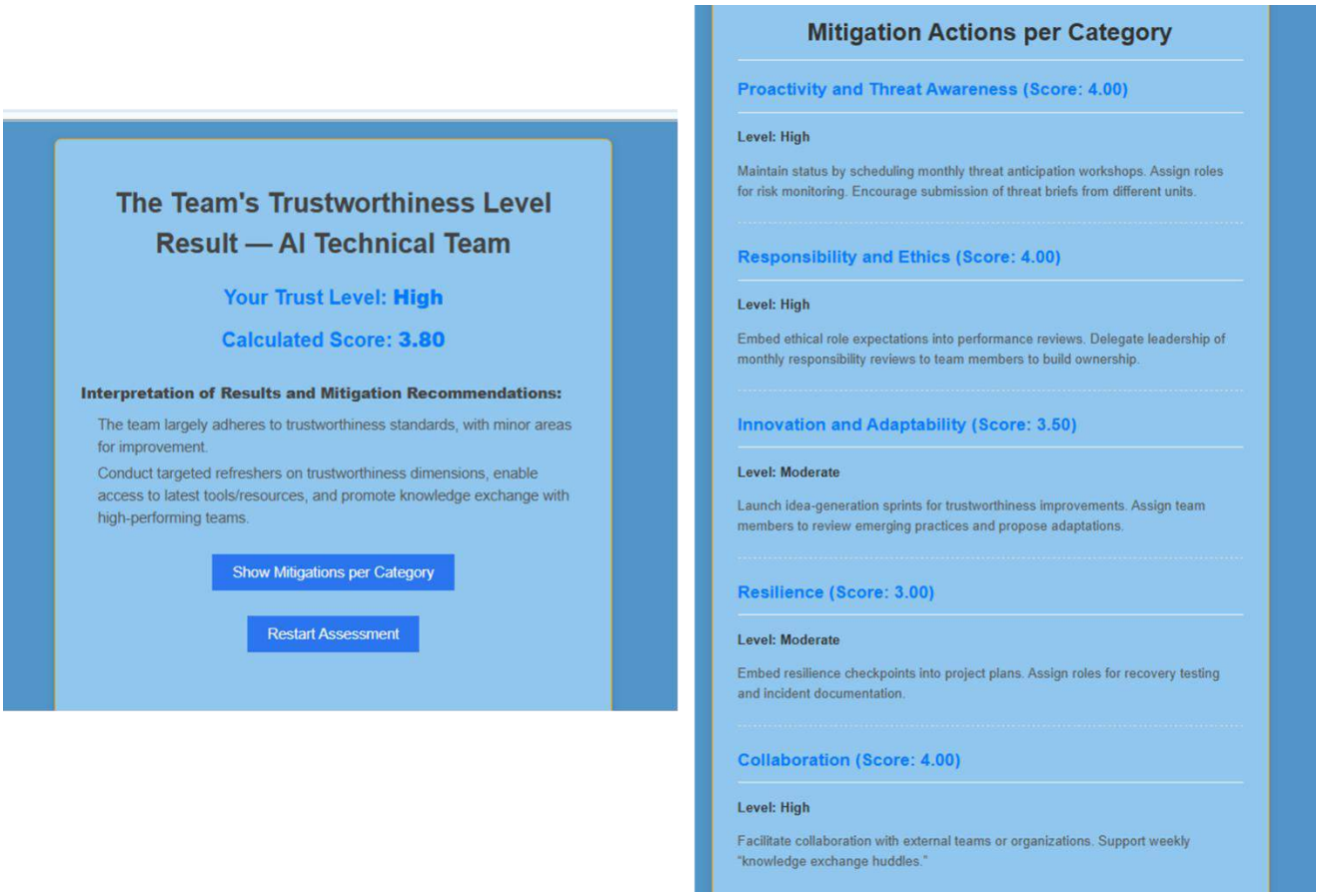


Figure 35: UNIFI trustSENSE results

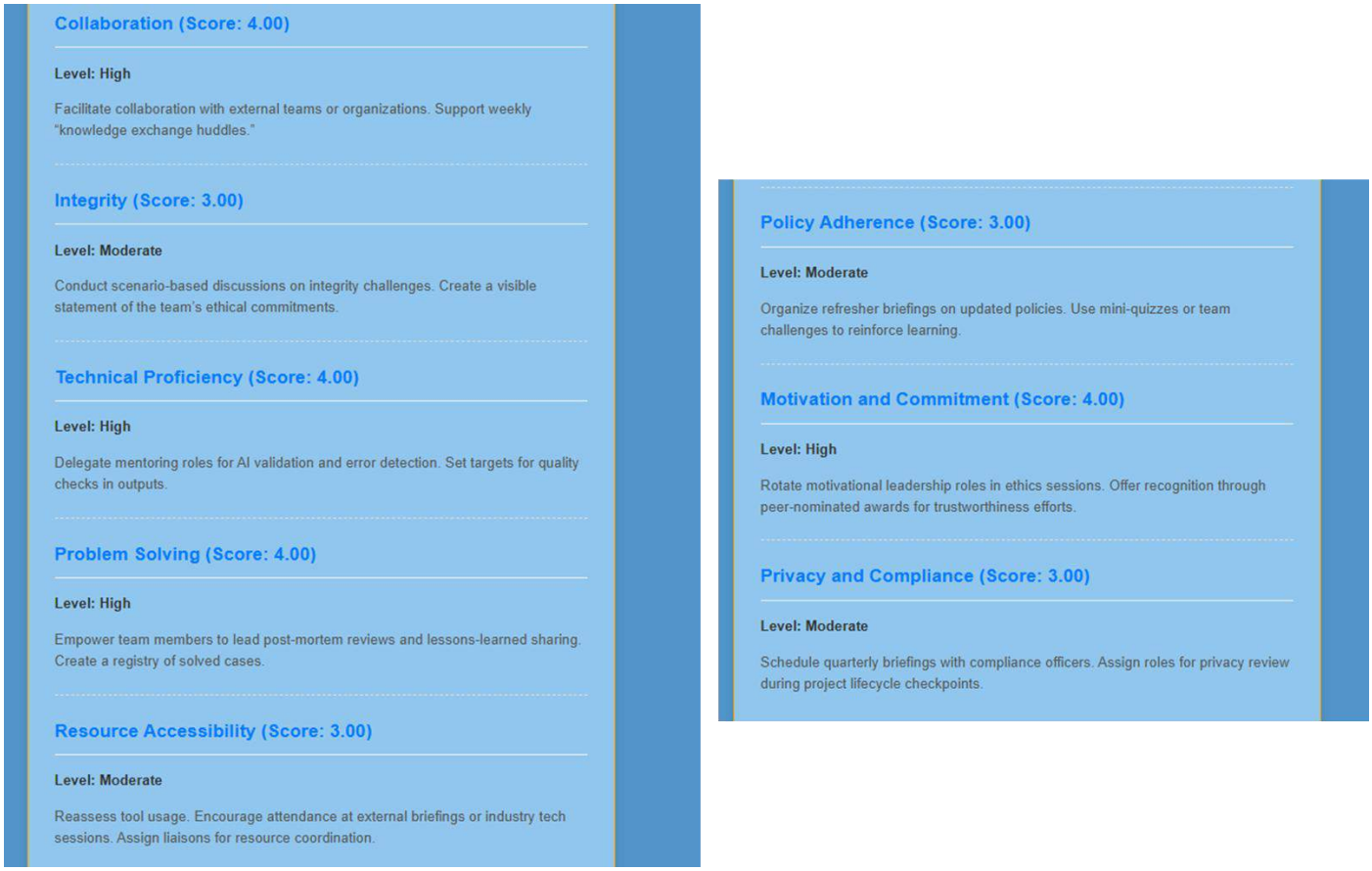


Figure 36: UNIFI trustSENSE results

3.3.2 Data & AI model Passport in the FAITH AI Model Hub

Data Passport

UNIFI used the FAITH AI Model Hub to create a Data Passport for the high-frequency FLS dataset used in the ATR pipeline (link). This is the same high-frequency sonar dataset described in Section 2.4, namely the dataset collected during the Ravenna pilot activities to support the close-range ATR workflow and the training, validation and testing of the high-frequency sonar branch. In this way, the data asset documented in Model Hub is directly connected to the pilot data-collection effort and to the current ATR implementation. The dataset registered by UNIFI is titled "High-Frequency FLS Seabed Inspection Training Dataset" and is identified by the acronym LSP4_HF_FLS. The registered data asset is clearly positioned within the specific operational context of the ATR pilot and within the refined sonar-based inspection logic adopted in Ravenna.

The authorship and ownership sections of the passport provide a clear governance context for the dataset. The contact person recorded for the asset is Alberto Topini, Assistant Professor, affiliated with the Department of Industrial Engineering, UNIFI. The publisher is listed as the FAITH Consortium, while the dataset owner is the Department of Industrial Engineering, UNIFI.

The motivation section further clarifies the intended scope of use of the dataset. The declared purpose is academic research, with non-commercial use only and research-specific restrictions. The passport explicitly identifies the suitable use of the asset as the training and validation of AI models for detailed sonar-based target instance segmentation, especially for close-range inspection and analysis of underwater targets in port environments. At the same time, the unsuitable uses are clearly documented: the dataset is not intended for autonomous critical decisions without human oversight, nor for use outside the validated sensing conditions without further testing, nor for commercial redistribution or third-party release without authorisation. This is an important contribution of the Data Passport, because it makes the permitted and non-permitted boundaries of the asset explicit and therefore supports more responsible reuse of the data.

The passport also records the dataset's role in the AI workflow. In the data-use section, the asset is associated with the Training phase and with Supervised Learning. In the distribution section, the dataset is described as consisting of high-frequency FLS acquisitions with instance-segmentation annotated polygons collected during the FAITH underwater robotics pilot. Access conditions are recorded as Restricted, with restricted rights and supporting links for documentation, direct access and data-dictionary resources. These elements are particularly useful for downstream trustworthiness analysis, since they make visible not only what the data are, but also how they were acquired, how they were annotated, and under which access regime they can be used.

Finally, the security and privacy section records dedicated governance measures for the dataset, including a Data Processing Agreement as legal measure, Usage Control as technical measure, and a Guideline as organisational measure, together with dedicated fields for personal-data characterisation. Even where the pilot is not centred on personal data, the presence of these fields is consistent with the FAITH objective of ensuring that data assets are documented with explicit attention to security, access control and compliance requirements.

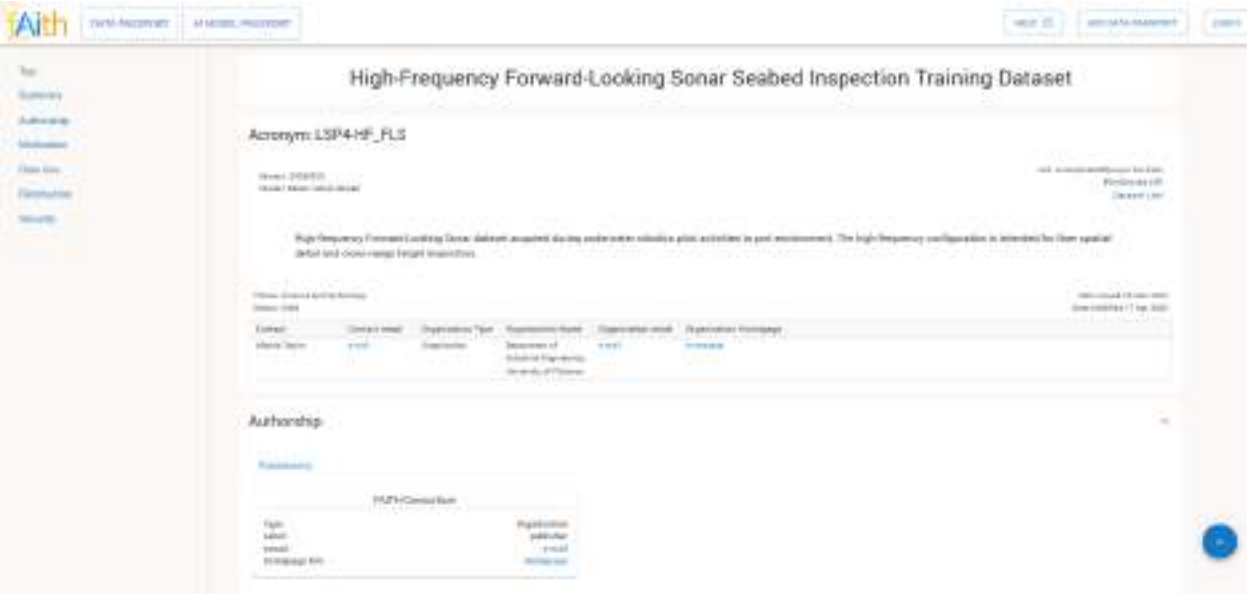


Figure 37: Snippet example from the Data Passport

AI Model Passport

UNIFI also used the FAITH AI Model Hub to create an AI Model Passport for the high-frequency FLS ATR model (link). It is important to underline that the documented model was trained on the corresponding high-frequency FLS training dataset described above, thereby preserving the explicit connection between the ATR model asset and its supporting training data.

The registered model is associated with the study "FAITH WP7 FLS Instance Segmentation Study". The passport explicitly links the model to the corresponding training dataset "High-Frequency FLS Seabed Inspection Training Dataset", thereby preserving the connection between the trained ATR model and the underlying sonar data asset on which the training process was based. The entry is associated with the Drones LSP, lists the Department of Industrial Engineering, UNIFI, as publisher, and identifies Fausto Fedi, UNIFI PhD Student, as author.

From a technical point of view, the passport specifies that the model is a YOLO11n-seg instance segmentation model, more precisely an Ultralytics YOLO segmentation mode. The model name recorded in the entry is "YOLO11n-seg FLS Instance Segmentation Model", the declared task is Instance Segmentation, and the learning method is Supervised Learning. The descriptive section of the passport states that the model is trained for the segmentation of FLS targets using a dataset configuration with four classes: ANVIL, MANNEQUIN, SEWER, and TIRE. It is intended for research, validation and pilot evaluation in port underwater environments, and not for unrestricted autonomous operational use. This information is particularly relevant because it makes the operational scope of the model explicit and directly connects the registered asset to the actual target categories used in the Ravenna pilot campaigns.

The passport also records explicit ethical concerns and a dedicated risk assessment. The ethical-concerns section notes that the model is intended for underwater target perception in a safety-relevant maritime context, and that incorrect segmentation or missed detections may negatively affect inspection or intervention decisions if the model is used without human oversight. For this reason, the model is documented as a decision-support tool to be used only within validated operational conditions. The risk assessment identifies the main risks as false positives, false negatives, reduced robustness outside the training domain, and overreliance on model outputs by operators. It further notes that additional risk may arise if the model is applied to sonar conditions, targets, or environments that are not represented in the training dataset.

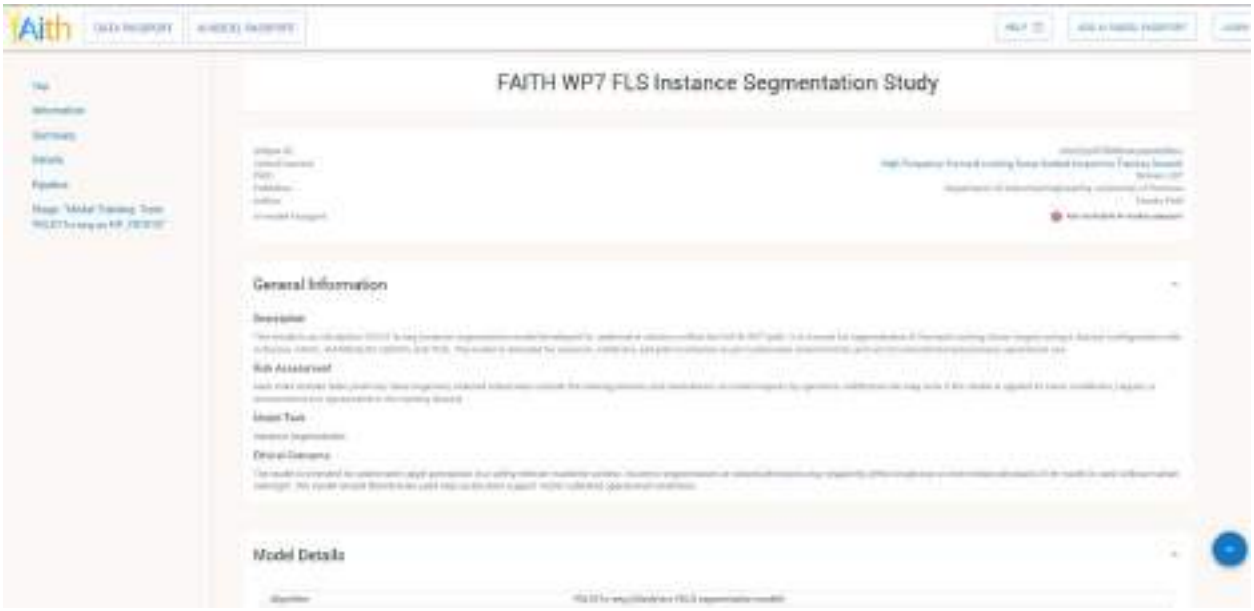


Figure 38: Snippet example from the AI Model Passport

For Replication Pilot

During the replication pilot, this activity will be extended further. In particular, UNIFI plans to include also the low-frequency FLS datasets and the corresponding low-frequency ATR models, to automate the passport-generation procedure through the Python library, and to progressively introduce into the FAITH AI Model Hub also the algorithms and area datasets associated with AC. In this way, the current high-frequency Data Passport and AI Model Passport should be regarded as the first step of a broader Model Hub integration process that will be expanded in the next pilot phase.

3.3.3 TrustGuard

UNIFI used TrustGuard to carry out the in-isolation assessment of two assets belonging to the underwater robotics pilot. The first was the *HF FLS dataset*, corresponding to the "*High Frequency*

FLS Seabed Inspection Training Dataset". The second was the trained ATR model built on that dataset, namely "*FAITH WP7 FLS Instance Segmentation Study*". In the current phase, the ATR dataset and ATR model were already documented through the FAITH AI Model Hub, while the AC assets had not yet been integrated into the AI Model Passport, and will be assessed in isolation through TrustGuard in the next months. In all two cases, the assets were positioned at PRODUCTION lifecycle stage and under High criticality, meaning that the assessments were performed on assets considered relevant for realistic pilot use rather than on purely hypothetical design-time artefacts.

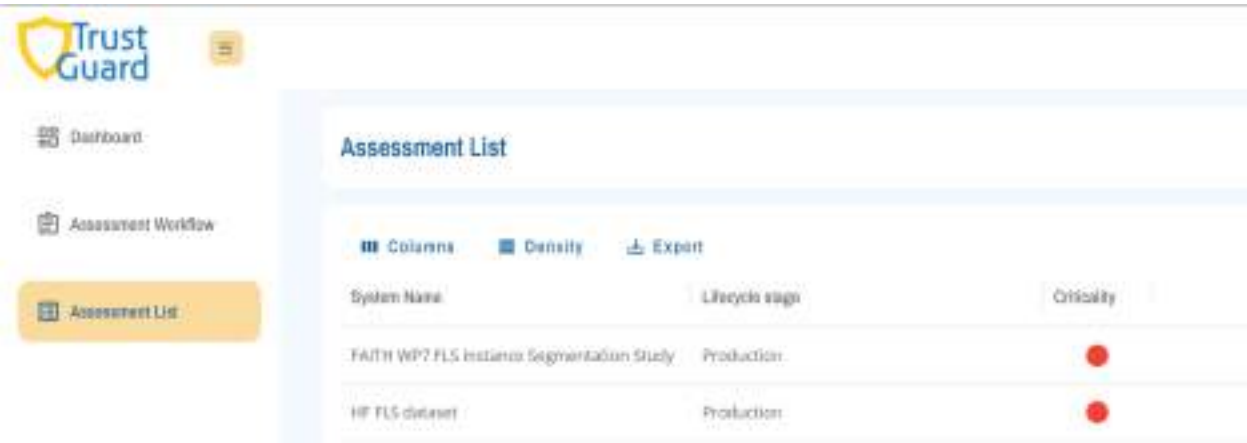


Figure 39: Visual snippet from TrustGuard with the two reported in-isolation assessments

For all two assets, the assessment activity started from the cartography and criticality stage, where UNIFI had to describe the asset under assessment, select the lifecycle stage, and complete the initial criticality questionnaire. The questionnaire asked whether the assessed AI system falls under the high-risk AI, system qualification of the AI Act, whether it is used by an Operator of Essential Services or for the provision of an essential or important service under NIS2, and what overall level of risk it may pose to health, safety or fundamental rights, including possible influence on decision-making.

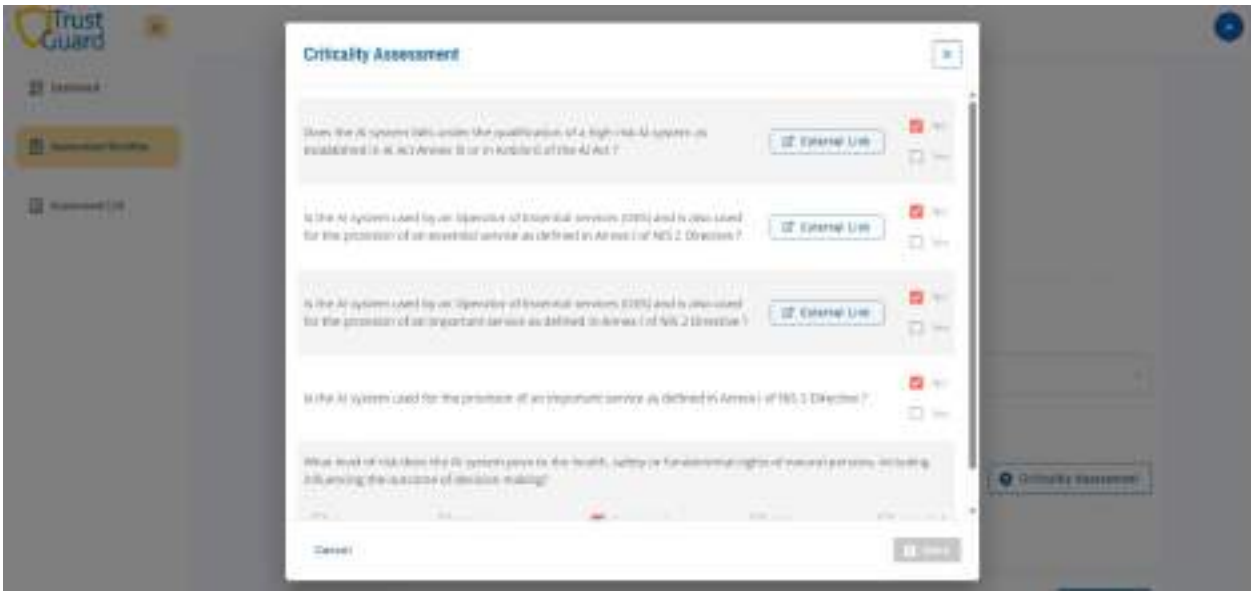


Figure 40: Example of TrustGuard Criticality Assessment

The core of the activity was then the completion of Section 2 – Assessment Questionnaire, which for all assets is structured into three main parts: Threat Assessment, Consequence Assessment, and Vulnerability Assessment. In the Threat Assessment part, UNIFI had to estimate how frequently a broad range of threats could occur in AI systems in the relevant domain. For the ATR assets, the questionnaire covered threats such as evasion, white-box or grey-box adversarial examples, oracle attacks, poisoning, label modification, model or data disclosure, data disclosure, model disclosure, compromise of ML application components, failure or malfunction of the ML application, human error, denial of service due to inconsistent data or sponge examples, cybersecurity incidents not reported, regulatory and legal non-compliance, reduced interpretability, model drift, lack of generalisability, poor data quality, limited diversity in training data, and changes in external conditions. In all cases, the questionnaire required a domain specific judgement on occurrence frequency using an ordinal scale ranging from once a year to never.

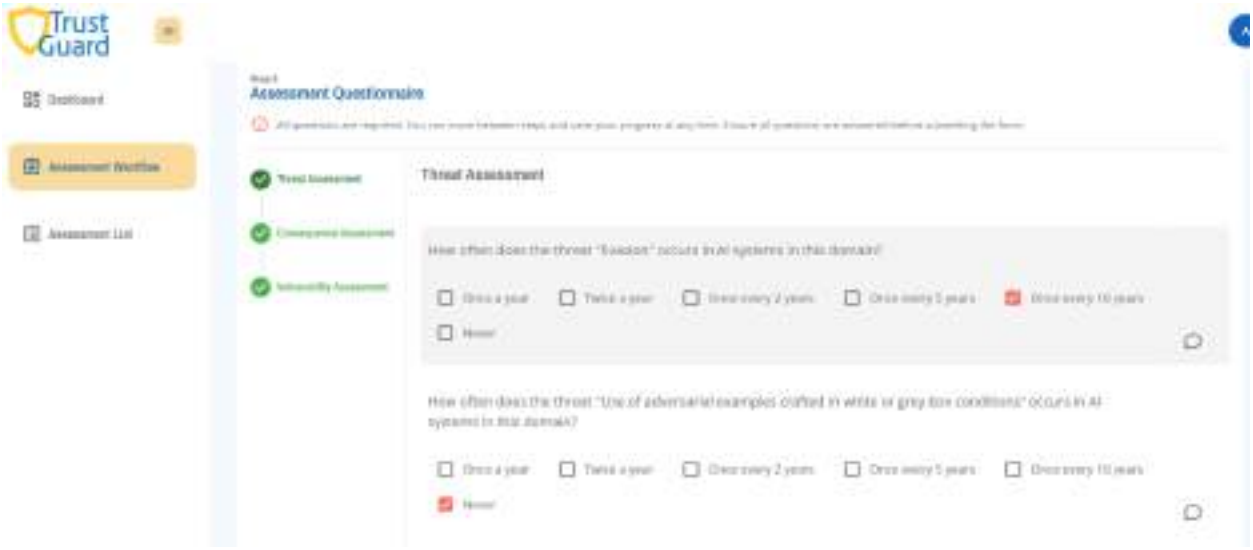


Figure 41: Example of TrustGuard Threat Assessment Questionnaire

After the threat frequency section, the questionnaire moved to the Consequence Assessment, where UNIFI had to evaluate the potential operational or business consequence associated with the compromise of the main trustworthiness properties. Across the assessed assets, the tool explicitly asked what the consequence level would be if the AI system's safe operation were compromised, if its security or resilience to adverse events were compromised, if the system were not interpretable or explainable, if there were a privacy compromise, if the system were not fair, if it were not transparent, and if it were not accurate or robust. The response scale ranged from Negligible to Very High. This part of the activity was important because it required UNIFI not only to think in terms of technical failure, but also to translate technical weaknesses into operational, business and trustworthiness consequences for the underwater robotics pilot.

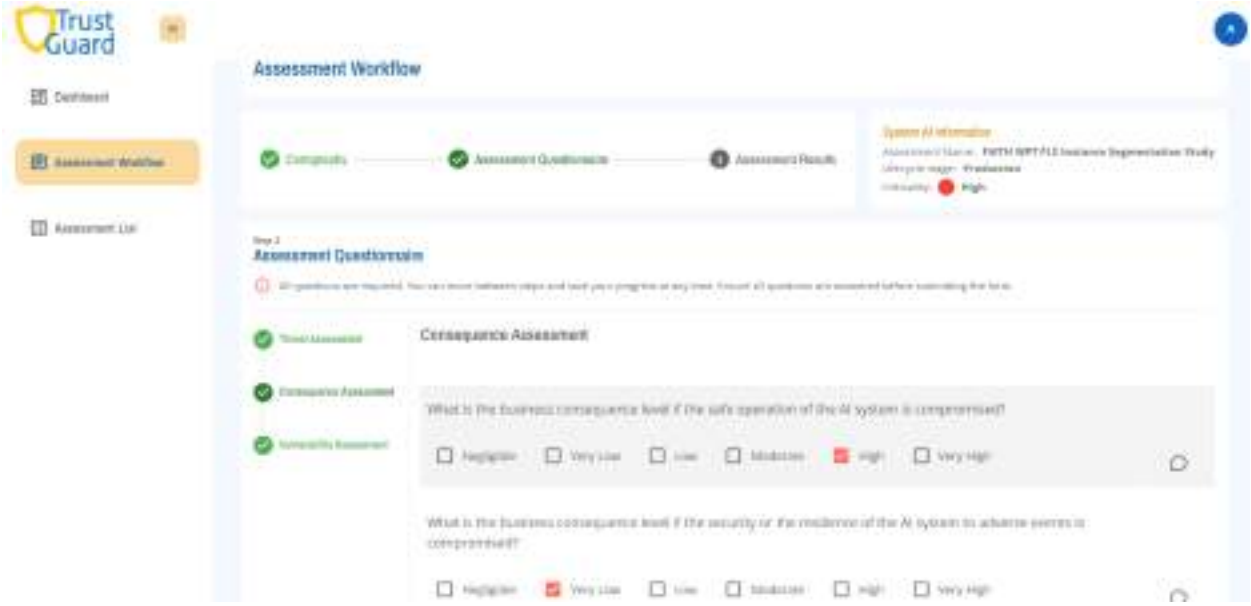


Figure 42: Example of TrustGuard Consequence Assessment Questionnaire

The third part of the questionnaire, the Vulnerability Assessment, was the most asset-specific section of the TrustGuard activity. For the ATR model, questions focused on the use of input, modification techniques to reduce exploitability, the existence of a security process for maintaining ML components and infrastructure, the evaluation of model exposure, the presence of adversarial or abnormal, data detection mechanisms, integration with incident, detection and response procedures, the use of adversarial training, limitation of publicly available information about the model, restriction of detailed outputs and architecture information, compliance with identity management and access, control policies, support for explainability and traceability, use of bias detection and mitigation methods, alignment of training and validation data with deployment conditions, compliance with applicable laws and regulations, and formal documentation standards for the model lifecycle.

For the HF FLS dataset, the more useful questions concentrated more strongly on dataset governance, provenance, security and representativeness. UNIFI had to answer whether adversarial detection or data, sanitisation techniques were applied to the training dataset, whether data, security controls were implemented across the dataset lifecycle, whether the dataset was diversified enough to improve robustness and generalisation, whether all data used for training, validation and inference were sourced through trusted and verifiable pipelines, whether external data sources were validated for trustworthiness and compliance, whether disclosure scenarios had been analysed, whether a strategy existed for ML, specific data attack vectors, whether privacy-preserving methods were used, whether provenance was systematically

tracked, whether legal or privacy experts had reviewed the data sources, and whether there was a structured policy for data-quality assessment during training and testing.



Figure 43: Example of TrustGuard Vulnerability Assessment Questionnaire

3.4. Threat Analysis

FAITH WP7 FLS Instance Segmentation Study

The TrustGuard results show that the ATR threat landscape is not dominated by a large number of equally important threats. On the contrary, the output is quite selective: only a small set of threats emerges as materially relevant, while most of the others remain low. This is important because it means that the model does not appear broadly exposed across the whole FAITH KB, but rather vulnerable in a few specific areas that matter directly for underwater deployment, namely component integrity, runtime reliability, robustness under changing conditions, and the adequacy of safeguards around model behaviour.

The threats that stand out most clearly in the tool output are the following:

- *Compromise of ML application components*. This is the strongest recurring threat in the results. It appears as Very High under Accountable and transparent and again as Very High under Safe. This means that compromise of the ML application stack is not treated by the tool as a narrow cybersecurity issue only. It is seen as a cross-cutting threat that can simultaneously undermine auditability, traceability, safety and operational trust in the ATR asset.

- *Failure or malfunction of ML application*. This is the second major threat family emerging from the results. It appears as High under Safe and as Very High under Valid and reliable. In practice, the tool is signalling that malfunction of the ATR application is one of the central threats to the asset because it directly affects whether the model can be relied upon in real deployment conditions.
- *Insufficient safeguards against bias*. This threat appears as High under Fair with harmful bias managed and rises to Very High under Explainable and interpretable. In the present use case, this should not be read narrowly as a social, fairness issue only. The way the tool scores it suggests a broader concern about whether the development and update lifecycle includes enough structured safeguards, checks and validation mechanisms to ensure that model behaviour remains understandable, stable and adequately governed.
- *Changes in external conditions*. This threat appears as High under Valid and reliable. This is a particularly meaningful result for ATR, because the model is intended for sonar-based underwater target perception, where environmental conditions, seabed appearance, acquisition conditions and scene variability can all affect performance. The tool therefore identifies deployment, condition shift as a concrete asset-level threat rather than as a purely theoretical issue.

Looking more closely at the structure of the results, the tool output suggests that the ATR threat profile is concentrated around three main problem areas.

The first is integrity of the *ML application stack*. The repeated emergence of compromise of ML application components at very high level means that the security and trustworthiness of the model cannot be separated from the security and trustworthiness of its surrounding software components, dependencies, and integration environment. In other words, the threat is not confined to the model weights or the inference engine alone; it concerns the whole operational chain that makes the ATR asset usable in practice.

The second is *runtime dependability*. The strong scoring of failure or malfunction of ML application under both safety and validity/reliability shows that the tool identifies malfunction as one of the most material threats to the ATR asset. This is especially relevant in the current pilot because the system is intended for underwater inspection support in a safety-relevant port context. A malfunction here does not merely imply degraded benchmark performance; it implies the risk of unreliable inspection support, missed targets, unstable behaviour or reduced confidence in the system during real operations.

The third is *insufficient lifecycle assurance*. The prominence of insufficient safeguards against bias and changes in external conditions suggests that the tool is capturing a more general issue: the model may be technically capable, but still insufficiently protected against performance drift,

update-time regressions, or deployment outside the immediate training and validation envelope. This means that the threat profile is not only about what the model does today, but also about how well its behaviour can remain controlled, interpretable and dependable as conditions evolve.

High Frequency FLS Seabed Inspection Training Dataset

The most important data, related threats identified by the tool are the following:

- *Poor quality of data used in training and testing*. Very High risk under Valid and reliable. This is one of the strongest dataset-specific results in the report and indicates that the main exposure of the HF FLS dataset lies in the possibility that the data may be insufficiently accurate, representative, or controlled for robust model training and evaluation.
- *Denial of service due to inconsistent data or a sponge example*. Very High risk under Valid and reliable, and High risk under Secure and resilient. For a dataset asset, this result points to the risk that unsuitable, malformed or inconsistent data may degrade downstream model behaviour or make the training and validation chain less dependable.
- *Regulatory and legal non-compliance*. High risk under Fair with harmful bias managed. This shows that, for the dataset, governance and compliance are not marginal issues. Provenance tracking, review of data sources and terms of service, and structured fairness-related checks emerge as relevant concerns in the assessment.

Taken together, these results show that the HF FLS dataset is not mainly threatened by high-end adversarial manipulation or by severe privacy leakage. Instead, the dataset-level threat profile is concentrated on whether the data are good enough, controlled enough, and governed well enough to support trustworthy model training and validation. In practical terms, the main issues raised by the tool concern quality of training/testing data, consistency of the dataset, and compliance/provenance, related governance, while disclosure, poisoning and privacy, related threats remain present but secondary.

3.5. Initial Risk Assessment results

FAITH WP7 FLS Instance Segmentation Study

The initial risk assessment shows that, for the current ATR asset, the overall risk profile is not uniformly high across all trustworthiness attributes. Instead, the TrustGuard output concentrates the highest risk levels in a small number of attributes, while most of the remaining risks stay at Low level. This is an important result, because it indicates that the model is not broadly critical in

every dimension, but is instead exposed through a few dominant weaknesses that are directly relevant for underwater deployment and operational use.

The highest risks are distributed as follows:

- *Safety*. This is one of the most critical attributes in the assessment. The threat "Compromise of ML application components" is rated Very High, while "Failure or malfunction of ML application" is rated High under the Safe property. This means that, from the tool's perspective, the main safety concern is not only wrong output generation in the abstract, but the possibility that weaknesses in the wider ML application stack or malfunction during operation may compromise the safe use of ATR in the pilot context.
- *Validity and reliability*. This is the other most critical attribute. Under Valid and reliable, the threat "Failure or malfunction of ML application" reaches Very High, and "Changes in external conditions" is assessed as High. This is a very significant result for the underwater robotics pilot, because it shows that the strongest residual risks are linked to dependable runtime behaviour and to the possibility that the model may be stressed by operating conditions that differ from those represented in the training and validation environment. In practical terms, this attribute captures the main technical concern for the ATR asset: whether the model remains reliable when deployed in real sonar-based underwater inspection.
- *Accountability and transparency*. Under Accountable and transparent, the threat "Compromise of ML application components" is again rated Very High. This shows that TrustGuard sees component compromise not only as a technical or security problem, but also as a direct challenge to auditability, traceability and trustworthy governance. If the broader ML application stack is not sufficiently controlled, the ability to justify, reconstruct and manage model behaviour becomes weaker, and this directly affects the transparency profile of the asset.
- *Explainability and interpretability*. Under Explainable and interpretable, the threat "Insufficient safeguards against bias" is rated Very High. In the current pilot context, this result should be interpreted broadly. It indicates that the model still lacks a sufficiently strong assurance layer to guarantee that its behaviour remains systematically interpretable and adequately controlled throughout updates and operational use. The tool is therefore signalling a major residual risk not only in fairness terms, but also in the structured verification of how the model's outputs can be understood and trusted.

Taken together, the tool output suggests a clear prioritisation of the trustworthiness attributes for the ATR asset. The highest, risk attributes are Validity and reliability, Safety, Accountability and transparency, and Explainability and interpretability. In practical terms, the risk assessment is

telling us that the main challenge for the current ATR asset is not whether it can be documented as a model, but whether it can remain technically dependable, operationally safe, traceable, and understandable once deployed in a real underwater robotics scenario.

When the human team is included in the picture, the overall assessment becomes more balanced. The trustSENSE output for the UNIFI AI Technical Team indicates an overall High trust level, with a calculated score of 3.80. This is important because it suggests that the human team is not the primary source of elevated risk in the current assessment. On the contrary, the team appears to provide a relatively strong organisational basis for managing the asset. The strongest maturity dimensions are Proactivity and Threat Awareness, Responsibility and Ethics, Collaboration, Technical Proficiency, Problem Solving, and Motivation and Commitment, all assessed at high level. This means that the technical team shows a solid capacity to develop, validate and improve the ATR system in a trustworthiness-oriented way.

At the same time, the trustSENSE result also shows several moderate dimensions, namely Innovation and Adaptability (3.50), Resilience (3.00), Integrity (3.00), Resource Accessibility (3.00), Policy Adherence (3.00), and Privacy and Compliance (3.00). These do not overturn the positive overall human-team result, but they are consistent with the areas where TrustGuard finds the main asset-level weaknesses. In particular, the moderate scores on resilience, policy adherence, and privacy/compliance align well with the fact that the highest model risks are concentrated in accountability, robustness, and lifecycle, governance dimensions. In other words, the combined reading of the two tools suggests that the team is technically strong, but that further formalisation of governance and resilience practices would help reduce the most important residual risks around the ATR asset.

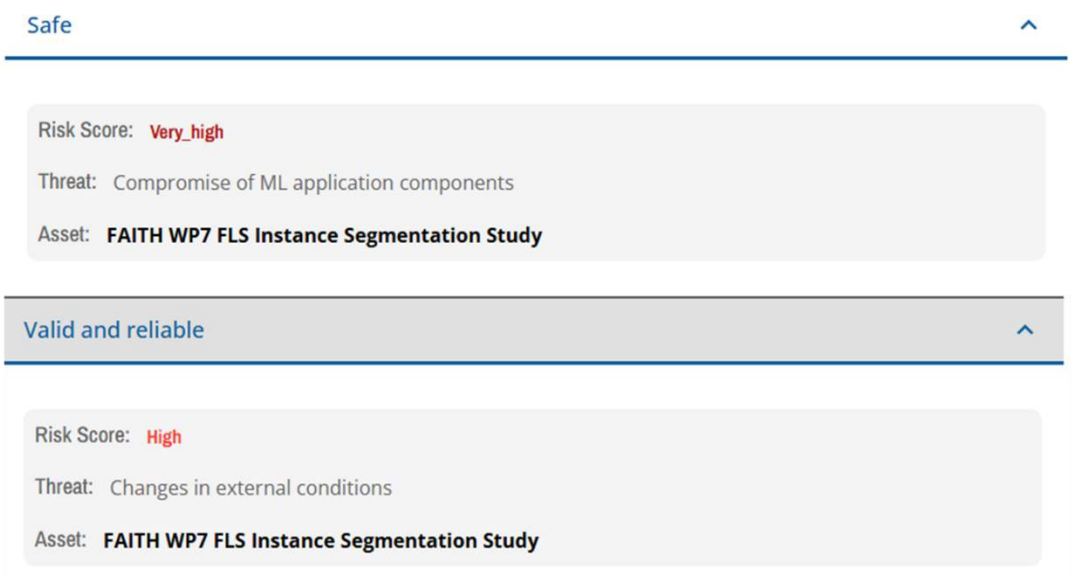


Figure 44: TrustGuard results for the FAITH WP7 FLS Instance Segmentation Study

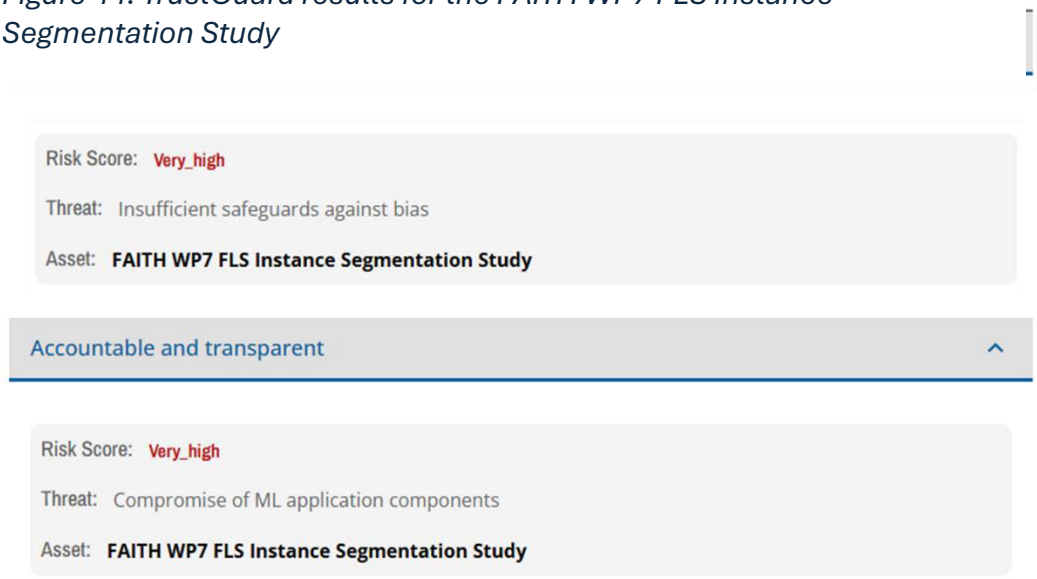


Figure 45: TrustGuard results for the FAITH WP7 FLS Instance Segmentation Study

High Frequency FLS Seabed Inspection Training Dataset

The strongest concentration of risk is associated with Validity and reliability, followed by Safety and Explainability and interpretability, while Privacy enhanced remains comparatively low.

The results by trustworthiness attribute can be summarised as follows:

- *Validity and reliable*. This is the attribute where the dataset shows the strongest concentration of severe risks. The threat "Failure or malfunction of ML application" is rated Very High, "Compromise of ML application components" is also Very High, "Denial of service due to inconsistent data or a sponge example" is Very High, and "Poor quality of data used in training and testing" is again Very High. For a data asset, this means that the dataset is primarily judged through its ability to support robust and dependable downstream model behaviour. The tool is therefore signalling that the most serious concern is whether the training and testing data are sufficiently reliable, consistent and fit for purpose to sustain a trustworthy ATR model in operation.
- *Safe*. The threat "Failure or malfunction of ML application" reaches Very High, while "Poisoning" remains Low. This indicates that the main safety, related issue for the dataset is not deliberate adversarial contamination, but rather the possibility that weaknesses in the data used for training and testing may contribute to unsafe downstream behaviour of the model. In other words, the dataset is seen as safety-relevant because poor or insufficiently representative data can later translate into unreliable model outputs in the operational environment.
- *Explainable and interpretable*. The threat "Failure or malfunction of ML application" is again rated Very High. For a dataset asset, this result suggests that weaknesses in the data can affect not only performance, but also the interpretability and trustworthiness of the resulting model behaviour. If the training data are not sufficiently aligned, controlled or representative, the downstream model may become harder to understand, justify or validate in practice.

Taken together, these results show a very clear pattern. For the HF FLS dataset, the highest risks are concentrated in the attributes that depend on the dataset's ability to support a robust, consistent and dependable ML lifecycle. The most critical attribute is Validity and reliability, followed by Safety, Explainability and interpretability. This means that the dataset is not primarily risky because it exposes sensitive information or because it is strongly vulnerable to adversarial manipulation; rather, it is risky because weaknesses in data quality, consistency, control and governance may propagate directly into the dependability of the trained ATR model.

In practical terms, the assessment is therefore telling us that the core trustworthiness issue for the HF FLS dataset is whether it is strong enough, controlled enough and representative enough

to sustain reliable downstream model behaviour. This is the central message emerging from the tool output and it explains why the highest, severity risks are clustered around validity, reliability and safety rather than around privacy or disclosure.

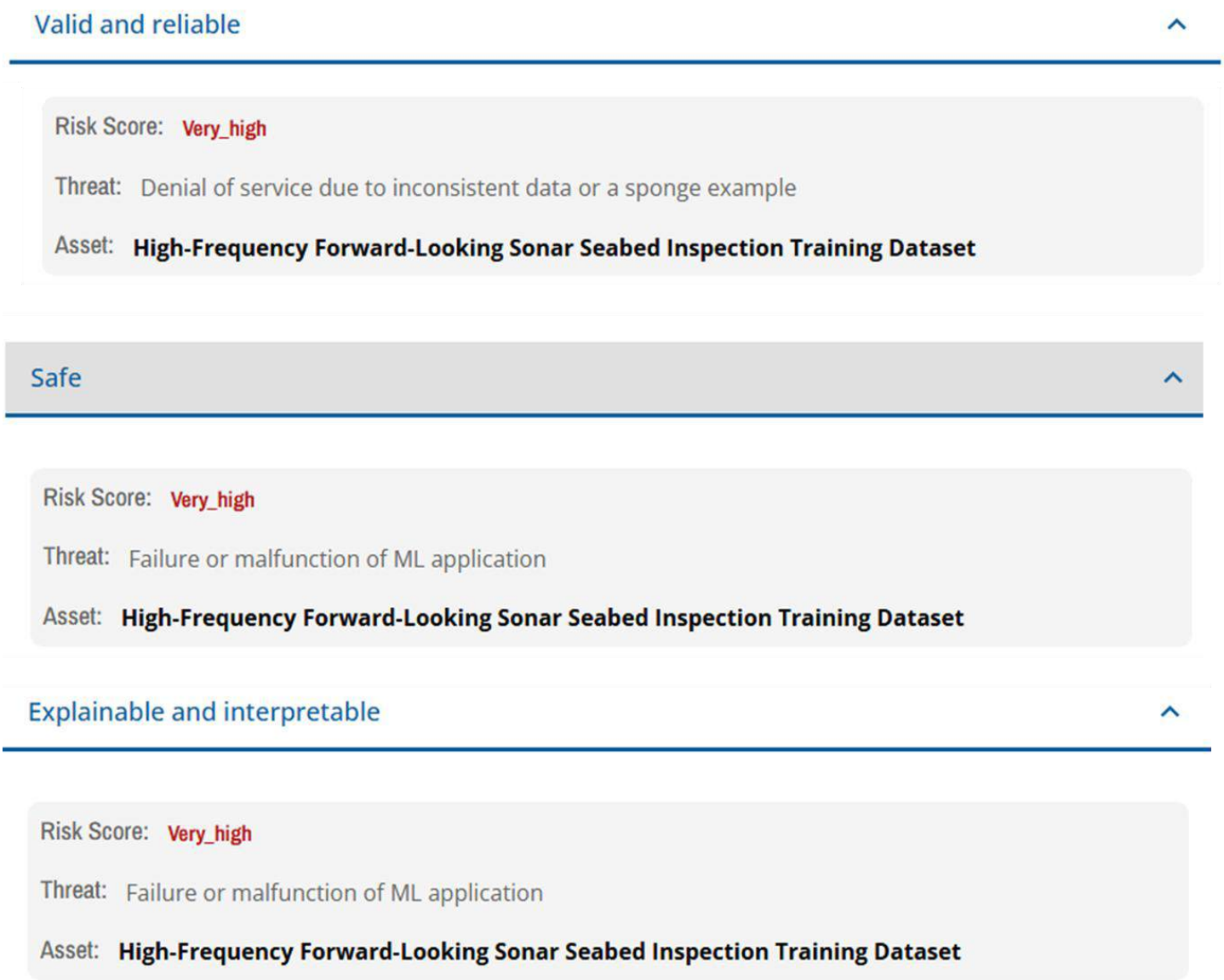


Figure 46: TrustGuard results for High Frequency FLS Seabed Inspection Training Dataset

3.6. Countermeasures

FAITH WP7 FLS Instance Segmentation Study

The recommended actions are strongly aligned with the highest, risk areas identified for the model, namely compromise of ML application components, failure or malfunction of the ML application, insufficient safeguards against bias, and changes in external conditions. The control set proposed by the tool is therefore centred on four practical intervention areas: security and component governance, runtime reliability and monitoring, robustness to deployment conditions, and structured update-time assurance.

The first intervention area concerns security and component governance of the ML application stack. For the highest, risk threat, namely compromise of ML application components, TrustGuard recommends controls such as reducing the information given by the model, integrating ML, specific concerns into existing security policies, checking the vulnerabilities of the components used, conducting a risk analysis of the ML application, ensuring compliance with security policies, and following the global process for integrating security into projects. In practical terms, UNIFI will implement some of these recommendations by formalising a dedicated ML application risk review for the ATR software chain, strengthening checks on dependencies, libraries and supporting components, and tightening control over which model artefacts, technical details and outputs are exposed outside the development environment.

The second intervention area concerns runtime reliability and malfunction prevention. For the threat failure or malfunction of ML application, TrustGuard recommends conducting a risk analysis of the ML application, ensuring that models are unbiased, including or integrating a drift, control service into the AI model, integrating ML applications into the overall cyber, resilience strategy, and again following the global security, integration process. In the LSP4 ATR context, these recommendations will be implemented by introducing a more structured model-control workflow. UNIFI will define and monitor clearer indicators of correct model functioning, including release-by-release performance checks, validation thresholds and failure monitoring on updated pilot data. In addition, the model will be re-validated after each significant update so that degradation in segmentation performance, instability across classes, or changes in prediction behaviour can be detected early.

A third intervention area concerns robustness to changing deployment conditions. For the threat changes in external conditions, the tool recommends employing adaptive or online learning techniques to handle shifting data distributions. In the underwater robotics scenario, this recommendation will be implemented in a proportionate way. Rather than introducing unconstrained online adaptation during operations, UNIFI will strengthen robustness by periodically re-training and fine-tuning the model on new high-frequency FLS pilot data,

expanding the diversity of the training material, and re-testing the model against the range of sonar conditions actually observed during deployments. This is particularly important because the model operates in a context where target appearance, acquisition geometry, background structure and environmental conditions can vary significantly. The implementation of this countermeasure therefore consists of a controlled update, and, revalidation loop, led by UNIFI-based on additional data curation and model update before later pilot deployments.

A fourth intervention area concerns structured safeguards during model updates, especially around explainability and interpretability. For the threat insufficient safeguards against bias, which reaches Very High under Explainable and interpretable and High under Fair with harmful bias managed, TrustGuard recommends regular bias regression tests during system updates, use of automated fairness testing tools within the model-development lifecycle, and embedding fairness criteria into CI/CD pipelines. In the present sonar-segmentation use case, these recommendations will not be interpreted in a purely demographic sense, but as the need for stronger systematic regression testing and release assurance. UNIFI will therefore implement per-release checks to verify that updates do not introduce undesired performance shifts across the target classes and operating conditions relevant to the pilot. In practice, this means moving toward a more formal update, validation discipline in which every new model version is compared against the previous one before deployment.

High Frequency FLS Seabed Inspection Training Dataset

For the HF FLS dataset, the first priority concerns data quality and consistency, because the strongest dataset-specific risks are linked to poor quality of data used in training and testing, denial of service due to inconsistent data, and failure or malfunction of the downstream ML application. In line with the tool recommendations, UNIFI will therefore implement a more structured data-quality assessment workflow for the HF FLS dataset. In practical terms, this means introducing explicit checks on annotation consistency, file integrity, completeness of metadata, class balance, acquisition quality, and compatibility of the samples with the intended training and testing tasks. The same action will also be supported by stricter control of all data used by the model, so that unsuitable, malformed, duplicated, or weakly documented samples are less likely to enter the training and validation pipeline. This action will be led by UNIFI, as owner and curator of the dataset, and will be carried out during the next improvement cycle of the ATR branch.

A second priority is dataset representativeness and robustness to operational conditions. The tool repeatedly recommends controls such as ensuring that the model is sufficiently resilient to the environment in which it will operate, enlarging the training dataset with more diverse data, and ensuring that training and test data reflect the actual operating conditions. For the HF FLS dataset,

this translates into a concrete action plan based on further expansion of the dataset with more diverse high-frequency sonar acquisitions, especially acquisitions covering different target poses, distances, seabed backgrounds, and environmental conditions observed during pilot activities. This activity will again be carried out by UNIFI, through additional curation of already collected material and the inclusion of new pilot acquisitions when available. The purpose is to reduce the gap between the current training dataset and the variability that may be encountered in real deployment. This action will proceed in parallel with the next rounds of dataset maintenance and model update.

A third priority concerns dataset governance and provenance control. Among the controls recommended by the tool are conducting a risk analysis of the ML application, tracking data provenance, ensuring reliable sources are used, and controlling all data used by the ML model. For the HF FLS dataset, this means that UNIFI will strengthen the governance layer around the dataset by improving traceability of data origin, clearer linking between raw acquisitions, annotations and derived splits, and more explicit documentation of which data are used for training, validation and testing. This action is particularly relevant because the TrustGuard assessment highlights regulatory and legal non- compliance as a high-risk issue for the dataset. In practical terms, this governance work will include version control of dataset releases, documentation of acquisition context, and a structured internal record of dataset updates and usage across the ATR lifecycle. Responsibility for this activity lies with UNIFI, with support from the project governance framework where compliance review is needed.

3.7 Formal Tool Feedback

The formal feedback activities for the FAITH tools were carried out in the second half of April 2026. The feedback was collected through the dedicated questionnaires provided for each tool, complemented, where relevant, by additional qualitative comments from the involved pilot partners.

TrustSense

The formal feedback process for trustSENSE was carried out through the completion of the dedicated questionnaire forms by UNIFI and through an additional written qualitative feedback note from the port-authority perspective (APRA and VPF). The questionnaire material covered both usefulness and usability, including ease of learning, clarity of interaction, support for human-related risk assessment, contribution to identifying human vulnerabilities, perceived fit with existing workflows, trust in the tool outputs, and intention for future use.

At a high level, the initial assessment was positive, in the sense that trustSENSE was considered useful for bringing the human dimension into trustworthiness assessment and for structuring

reflection on organisational preparedness, user roles and stakeholder alignment. At the same time, the qualitative feedback highlighted an important limitation for early-stage adoption contexts: several questions assume that AI systems are already fully integrated into day-to-day organisational practice, with established governance structures, audits, quantified performance indicators and AI-specific responsibilities. In pilot contexts, this may bias results toward lower scores even when the organisation is generally mature and well prepared, simply because the relevant AI-specific processes are not yet formally in place.

The main improvement areas identified concern the need for a more context-sensitive and maturity-adaptive questionnaire, clearer differentiation between lack of capability and lack of responsibility in multi-actor settings, reduced dependence on rigid quantitative thresholds, and the introduction of a "Not Applicable" option for questions that do not fit the current stage of adoption. Overall, the formal feedback confirms that trustSENSE is already useful as an awareness-raising and structuring tool, while its main refinement need concerns better applicability to pilot-phase and distributed-responsibility environments.

Data & AI model Passport in the FAITH AI Model Hub

The formal feedback process for the Data Passport and AI Model Passport functionalities of the FAITH AI Model Hub was carried out by UNIFI through the completion of the dedicated questionnaire forms provided for the tool. The questionnaires addressed both usefulness and usability, including ease of learning, clarity of interaction, usefulness for documentation and risk-assessment work, support for transparency and explainability requirements, and overall ease of use.

At a high level, the initial assessment was positive. UNIFI considered the AI Model Hub useful as a structured environment for documenting datasets and models, storing metadata, and improving traceability and transparency of AI assets. This was particularly relevant in the pilot because the high-frequency FLS dataset and the corresponding ATR model were documented through these functionalities.

The main usability improvement highlighted by UNIFI concerns the time required to register new assets, especially when users need to insert assets that are very similar to existing ones. In particular, it was suggested that the AI Model Hub should support a duplication function for datasets and models, so that users can copy an existing asset and modify only the necessary fields. This would make the platform more user-oriented, reduce effort, and limit the risk of inconsistencies during data entry

TrustGuard

The formal feedback process for TrustGuard was carried out by UNIFI through the completion of the dedicated questionnaire forms provided for the tool. The feedback covered both usefulness and usability, including ease of use, clarity of interaction, support for documentation, and usefulness for identifying threats, vulnerabilities and trustworthiness risks in in-isolation asset assessments.

At a high level, the initial assessment was positive. UNIFI considered TrustGuard useful as a structured tool for producing more consistent risk-assessment outputs and for guiding the analysis of AI assets through a clear workflow. The main improvement areas identified concern more asset-specific questionnaires, the possibility to duplicate an existing assessment for similar or evolving assets, and a more user-friendly and streamlined interface. Overall, the feedback confirms that TrustGuard is already useful for structured risk assessment, while its main refinement needs concern adaptability, efficiency and usability.

4. AI System Evaluation

4.1. Quantitative metrics on pilot data

ATR – High-Frequency

The following ATR quantitative evaluation currently refers to the high-frequency FLS model trained for instance segmentation on four target classes, namely ANVIL, MANNEQUIN, SEWER and TIRE. The training configuration is based on a YOLO-based segmentation pipeline specifically adapted to the high-frequency sonar branch and fine-tuned on pilot data collected in Ravenna. In line with the KPI framework defined for ATR, the evaluation focuses on detection performance through precision, recall, mean Average Precision, false positive rate and false negative rate, as well as on operational efficiency and robustness in low-visibility underwater conditions. The D7.1 KPI framework sets the target direction for ATR at high precision and recall, high mAP, low false positive and false negative rates, near-real-time processing, throughput above 5 fps, and limited degradation in murky conditions.

The first set of results reported here was obtained by taking the data collected during the first three experimental campaigns, and splitting the resulting high-frequency FLS dataset into training, validation and test subsets. The quantitative metrics reported in this first evaluation therefore refer to the test split only, meaning that they were computed on images never seen during training or validation. This is an important point, because it shows that the reported values do not measure memorisation of the training material, but actual generalisation to unseen pilot data acquired within the first part of the Ravenna campaign sequence.

On this unseen test split, the high-frequency FLS ATR model showed a very strong performance profile. At mask level, the overall mAP@0.5 is 0.982, while the best overall F1 score is about 0.96 at a confidence threshold around 0.53. Precision remains very high across a broad confidence range and reaches 1.00 at high confidence, while recall stays close to 0.98–0.99 over most of the operating interval and drops only at very high thresholds. Class-wise behaviour is also strong: MANNEQUIN is the best-performing class, while SEWER is comparatively the most difficult one, but still remains at a very high level.

Table 3: Performance of the high-frequency FLS ATR model on the unseen test split derived from the first three experimental campaigns

KPI	Result
Precision	95.6%
Recall	97.6%
F1-score	96.0%
mAP@0.5	98.2%
mAP@0.5:0.95	63.0%
mAP@0.5 – ANVIL	98.6%
mAP@0.5 – MANNEQUIN	98.7%
mAP@0.5 – SEWER	96.7%
mAP@0.5 – TIRE	98.6%
False Positive Rate	4.4%
False Negative Rate	2.4%
Sensor acquisition rate	3 Hz

KPI	Result
Model inference throughput on NVIDIA AGX Orin	~8 Hz
Detection latency	~0.125 s/frame
Optical effective visibility in pilot conditions	~0.5 m
Effective target visibility with high-frequency sonar + ATR	up to ~5 m

These results are fully coherent with the ATR KPIs defined in D7.1. Precision, recall and mAP are all above the expected target range, the estimated FPR and FNR remain low, and inference throughput is well above the sonar acquisition rate. In practical terms, the ATR node running on the NVIDIA AGX Orin onboard FeelHippo AUV is able to process the incoming sonar stream in near-real time, which means that latency and throughput requirements are satisfied. Moreover, the results are especially relevant in relation to the low-visibility KPI: while optical cameras reached a practical visibility of around 0.5 m, the high-frequency sonar combined with the ATR model allowed target observation and recognition up to around 5 m, thus preserving operational perception capability precisely in the environmental conditions where optical sensing becomes severely limited.

In addition to this standard test-set evaluation, UNIFI also carried out a dedicated robustness-oriented check using data from the last experimental campaign. In this campaign, all targets had been positioned differently on the seabed with respect to the earlier acquisitions, creating a more challenging distribution shift. For example, the mannequin was lying on the bottom in a different pose, and more generally the geometric presentation of the targets differed from the configurations most represented in the original train/validation/test split. The purpose of this additional assessment was to verify how robust the model remained when exposed to target arrangements that were still operationally meaningful, but less similar to those on which it had been trained and validated.

As expected, the model remained clearly usable, but performance degraded with respect to the standard test split. The degradation was intentionally more visible on the false positive side, meaning that the model preserved a relatively good recall but became more permissive in the presence of unfamiliar target poses and changed background/echo configurations. This is consistent with the expected behaviour of a perception model exposed to a mild domain shift: true targets are still often detected, but the altered target geometry and scene appearance

increase the probability of activating detections on ambiguous structures or on target-like acoustic patterns.

Table 4: ATR high-frequency FLS model: robustness check on the last experimental campaign with differently positioned targets

KPI	Robustness, check result
Precision	88.9%
Recall	93.1%
F1- score	90.9%
mAP@0.5	92.2%
mAP@0.5:0.95	54.8%
False Positive Rate	12.8%
False Negative Rate	6.9%
Sensor acquisition rate	3 Hz
Model inference throughput on NVIDIA AGX Orin	~8 Hz
Detection latency	~0.125 s/frame
Effective target visibility with high-frequency sonar + ATR	up to ~5 m

From an interpretative point of view, the comparison between the two evaluations is informative. On the one hand, the standard unseen test split built from the first three campaigns demonstrates that the high-frequency FLS ATR model already reaches a high level of technical maturity, with very strong precision, recall and mAP values, low error rates, and real-time compatibility with onboard deployment. On the other hand, the additional robustness check confirms that the model is still sensitive to distribution changes in target pose and scene configuration, and that this sensitivity manifests itself mainly through an increase in false positives rather than through a catastrophic collapse in recall. This is an important result for the pilot, because it highlights the practical value of continuing the robustness-oriented strategy already foreseen in the project, namely iterative dataset extension, retraining with more heterogeneous target presentations,

and the future integration of safeguards such as OOD-aware monitoring and stronger operator-facing verification mechanisms.

As a whole, the current quantitative evidence for the high-frequency FLS ATR model is very positive. The model satisfies the core D7.1 detection-performance and operational-efficiency KPIs on the standard unseen test split, while the additional last-campaign robustness check confirms that the main residual weakness is not a loss of basic recognition capability, but a more noticeable increase in false positives under changed target-placement conditions. In this sense, the initial pilot already demonstrates strong ATR performance on pilot data, while also providing a clear direction for the next improvement cycle focused on generalisation and reduction of background-driven false alarms.

ATR – Low-Frequency

The same quantitative evaluation logic was also applied to the low-frequency FLS ATR model, which represents the longer-range branch of the sonar-based recognition pipeline. While the high-frequency model was designed for closer and more detailed inspection, the low-frequency model was configured to support earlier target discovery at longer stand-off distances, with an effective operational range of up to approximately 10 m. As in the high-frequency case, the model runs onboard the NVIDIA AGX Orin and reaches an inference rate of approximately 8 Hz, while the sonar acquisition rate during pilot operations was approximately 3 Hz. Therefore, also for the low-frequency branch, the processing pipeline remains compatible with near-real-time onboard use.

For this model, the current quantitative evaluation refers to the subset of target classes that remained meaningful in the low-frequency branch, namely MANNEQUIN, SEWER, and TIRE. The first evaluation was performed by taking the data collected during the first three experimental campaigns, and splitting the resulting low-frequency dataset into training, validation and test subsets. As for the high-frequency model, the results reported below refer to the test set only, that is, to images never seen during training or validation.

On this unseen test split, the low-frequency model showed a solid performance profile, although lower than the corresponding high-frequency branch, which is expected given the coarser acoustic representation and the lower amount of discriminative target detail available at longer range. At mask level, the overall mAP@0.5 is 0.932, with class-wise values of 0.933 for MANNEQUIN, 0.959 for SEWER, and 0.903 for TIRE. The corresponding best overall mask F1-score is 0.92, reached at a confidence threshold of approximately 0.515. At box level, the overall mAP@0.5 is 0.965, with a best overall box F1-score of 0.94 at a confidence threshold around 0.392. The curves show that precision remains high across a broad range of thresholds, while recall degrades more quickly than in the high-frequency branch as the confidence threshold

increases. This is coherent with the fact that low-frequency sonar provides a wider operational envelope, but also produces less detailed and less distinctive target features, making fine-grained separation from background structures intrinsically harder.

The model still separates the three target classes reasonably well, with limited direct inter-class confusion. Most residual errors are instead associated with background-related activations and missed detections, rather than with systematic swapping of one class for another. In other words, when the model fails, it more often fails by missing a target or by generating additional detections on ambiguous background returns, rather than by consistently confusing MANNEQUIN, SEWER and TIRE with one another. This behaviour is important because it indicates that the main limitation of the low-frequency branch is not class discrimination per se, but the weaker discriminative quality of the underlying features

Table 5: ATR low-frequency FLS model: evaluation on unseen test split derived from the first three experimental campaigns

KPI	**Result**
Precision	93.6%
Recall	92.4%
F1 – score	92.0%
mAP@0.5	93.2%
mAP@0.5:0.95	45.0%
mAP@0.5 class – MANNEQUIN	93.3%
mAP@0.5 class – SEWER	95.9%
mAP@0.5 class – TIRE	90.3%
False Positive Rate	9.0%
False Negative Rate	7.5%
Sensor acquisition rate	3 Hz

KPI	Result
Model inference throughput on NVIDIA AGX Orin	~8 Hz
Detection latency	~0.125 s/frame
Effective target visibility with low-frequency sonar + ATR	up to ~10 m

Precision and recall remain high, mAP is still strong across the three retained classes, and the inference rate remains significantly above the sonar acquisition rate. However, compared with the high-frequency model, the low-frequency branch shows a more evident trade-off between range and discriminative quality: it sees farther, but the target representation is less rich, and this makes the model more prone to borderline detections and background-driven uncertainty. In addition to the standard test-set evaluation, UNIFI also carried out a dedicated robustness-oriented stress test using data from the last experimental campaign, where the targets had been repositioned differently on the seabed. This created a more challenging distribution shift than the one already observed for the high-frequency branch. The purpose of this test was to verify how the low-frequency model generalises when target configurations change.

As expected, performance degraded more noticeably than in the high-frequency case. The reduction was in the order of approximately 8–12%, and the dominant effect was again an increase in false positives, but in this case more accentuated. This stronger degradation is coherent with the nature of the low-frequency sensing regime. Because the low-frequency branch works with poorer and less detailed target features, the learned representation is inherently less discriminative and the model finds it more difficult to generalise when the same objects are seen under significantly different poses or bottom-contact configurations. In other words, when the target geometry changes, the lower-resolution sonar signature becomes more ambiguous and background structures are more easily interpreted as plausible detections.

Table 6: ATR low-frequency FLS model: robustness check on the last experimental campaign with differently positioned targets

KPI	Robustness, check result
Precision	83.8%
Recall	84.9%

KPI	Robustness, check result
F1-score	84.3%
mAP@0.5	84.5%
mAP@0.5:0.95	35.0%
Precision	87.2%
Recall	86.0%
False Positive Rate	19.5%
False Negative Rate	13.0%
Sensor acquisition rate	3 Hz
Model inference throughput on NVIDIA AGX Orin	~8 Hz
Detection latency	~0.125 s/frame
Effective target visibility with low-frequency sonar + ATR	up to ~10 m

From an interpretative point of view, the comparison between the two low-frequency evaluations is quite meaningful. On the one hand, the standard unseen test split built from the first three campaigns shows that the low-frequency model already provides a useful and operationally relevant long-range detection capability, with solid mAP, good recall, and processing speed compatible with real-time deployment. On the other hand, the robustness test on the last campaign shows that the model is more sensitive than the high-frequency branch to target repositioning and feature variation, and that this sensitivity manifests itself mainly through higher false positive activity. This is the expected price of working at longer range with a coarser signal representation: the system sees farther, but with weaker local detail and therefore lower generalisation strength.

Overall, the low-frequency branch should therefore be interpreted as a search-oriented ATR layer, particularly valuable for extending the observation range up to around 10 m, while accepting a somewhat lower robustness than the high-frequency close-inspection model. In operational terms, this reinforces the rationale of the dual-sonar ATR workflow adopted in the pilot: the low-

frequency model is well suited for earlier discovery over a wider area, whereas the high-frequency model provides the stronger and more reliable confirmation stage once the vehicle moves closer to the candidate target.

AC

For the AC branch, the quantitative evaluation was carried out through field validation on a dataset of distinct survey areas. The algorithm was tested on 20 distinct validation areas, each defined as a polygonal region of interest in the port environment. Taken together, these areas represent approximately 26,048 m² of surveyed surface, with an average area of about 1,302 m², a median area of about 1,393 m², and a size range between roughly 389 m² and 2,069 m². This provides a sufficiently varied validation set for assessing the practical behaviour of the planner on areas of different shape and extension.

Table 7: AC validation areas used for field testing

Validation, set indicator	**Value**
Number of validation areas	20
Total surveyed area	~26,048 m²
Mean area per validation polygon	~1,302 m²
Median area	~1,393 m²
Minimum area	~389 m²
Maximum area	~2,069 m²

The quantitative evaluation of AC was then aligned with the KPI framework defined in D7.1, namely coverage completeness, survey duration, data resolution, mission success rate, real-time processing latency, operational cost efficiency, and safety improvement. In the current pilot, the most robust evidence concerns coverage completeness, mission duration, mission success, real-time planner responsiveness, and safety, while operational cost efficiency can at this stage only be estimated preliminarily.

The strongest result is related to coverage completeness. In all the validation cases, the AC algorithm was able to reach the target coverage percentage defined for the mission. In other

words, the planner successfully generated and executed trajectories that satisfied the requested coverage level in 100% of the tested cases. Since the D7.1 KPI requires optimisation of routing to achieve at least 90% coverage, this target can be considered satisfied in the current field validation. This is an important result because it confirms that the AC pipeline is already able to provide dependable area coverage under real pilot conditions, rather than only in simulation. The second key quantitative dimension is survey duration.

In the pilot, the survey duration of the autonomous missions was measured directly during field operations, taking into account the actual execution of the coverage task in the validation areas. For representative areas of the scale considered in the current pilot, the measured autonomous mission duration was generally in the order of 30–40 minutes, including the practical execution of the survey and the associated mission-management phases. By contrast, the duration of an equivalent manned survey solution was not measured directly in the pilot, but estimated conservatively on the basis of a realistic operational workflow. This estimate included boat preparation, crew mobilisation, launch or undocking, transit to the inspection area, execution of the survey, and return logistics. For areas of comparable scale, the corresponding duration was estimated at approximately 55–65 minutes. Based on this comparison, the AC workflow shows a clear reduction in overall mission time with respect to a conventional manned procedure, comfortably exceeding the D7.1 target of at least 10% reduction. Even under conservative assumptions, the time saving remains substantial. This is an important operational result, because it shows that the autonomous solution is not only effective in terms of coverage, but also competitive in terms of total mission duration once realistic field logistics are taken into account.

The data-resolution KPI must be interpreted in light of the multibeam sensor actually used in the pilot. The AC system employed a Norbit WideBand Multibeam Sonar, which provides 256–512 beams, 400 kHz nominal operating frequency with frequency agility up to 700 kHz, an adaptive ping rate up to 60 Hz, and depth resolution below 10 mm. These characteristics indicate that the sensor is capable of producing dense bathymetric point clouds and detailed local seabed reconstruction, with the final point density depending on altitude, selected swath sector, overlap, and vehicle speed. In the present pilot, the combination of this multibeam configuration with a vehicle speed of 0.5 m/s proved sufficient to generate bathymetric outputs with enough spatial detail to support coverage monitoring and practical seabed reconstruction. Therefore, while an exact point-density-per-square-metre value was not formally extracted in this initial pilot phase, the achieved bathymetric quality is consistent with the KPI objective of maintaining high data density without sacrificing operational usability.

A more critical result concerns the mission success rate. Out of 20 executed AC missions, 4 missions showed technical problems related primarily to the innovative and still prototypal nature of the vehicle platform, rather than to the coverage-planning logic itself. This leads to a current

mission success rate of 80%. This value is below the D7.1 target of over 95%, and therefore this KPI cannot yet be considered achieved. The result is nevertheless informative, because it indicates that the main limitation at this stage is not the conceptual validity of the coverage algorithm, but the current stability of the overall robotic system. For this reason, vehicle stabilisation and integration hardening will be one of the priorities before the replication pilot in Valencia.

The real-time processing latency result is positive. During execution, the planner started computing the next planning segment before the vehicle reached the next waypoint, so that the updated route was already available and could be applied immediately upon arrival. From an operational point of view, this means that the planning and replanning cycle did not introduce perceptible waiting times between waypoints. Even if an exact latency value was not logged as a standalone benchmark, the observed behaviour is consistent with the KPI requirement of maintaining planner responsiveness within a few seconds, and practically corresponds to a real-time-ready planning loop.

The safety improvement KPI is also strongly supported by pilot evidence. No diver was involved in the execution of the AC missions, and no operator was required onboard a support vessel during the survey itself (except for the failed missions). The only human intervention foreseen in the current operational setup is the human-on-the-loop strategy, in which an operator supervises the mission from the quay and can interact with or take control of the vehicle if needed. In this sense, the AC system already demonstrates one of its most relevant trustworthiness benefits: it substantially reduces direct human exposure to underwater hazards and removes the need for diver-based intervention during routine coverage operations. Within the current pilot, no safety incident involving human operators was recorded, so the safety-improvement trend is clearly positive.

A preliminary remark can also be made on operational cost efficiency. Although a formal cost-per-square-metre accounting has not yet been completed in the initial pilot, the observed deployment logic already indicates a favourable trend: the absence of diver deployment and the absence of a dedicated manned survey vessel during mission execution reduce personnel, logistics and operational overhead. A more formal cost comparison will be appropriate in the replication phase, when the system stabilisation effort will make repeated operational measurements more comparable.

Table 8: AC quantitative metrics on pilot data

KPI	Result
Coverage completeness	Target coverage reached in 100% of tested validation areas
Survey duration	~30–40 min autonomous mission for representative areas of current scale
Estimated equivalent manned procedure	~55–65 min including mobilisation and boat-related logistics
Mission time reduction	~30–40% estimated reduction
Data resolution	High-density bathymetric acquisition enabled by 256–512 beam multibeam sonar, adaptive ping rate up to 60 Hz, and <10 mm depth resolution
Mission success rate	16/20 missions = 80%
Real-time processing latency	Planner updated trajectory before arrival at next waypoint; no operational waiting observed
Operational cost efficiency	Positive preliminary trend; no formal €/m² evaluation yet
Safety improvement	No diver involved, no onboard survey operator required, shore-based human-on-the-loop only

Finally, the algorithm demonstrated full success in meeting the requested coverage percentage across all tested validation areas, showed a clear reduction in mission duration with respect to a realistic manned procedure, delivered bathymetric data of sufficient quality for practical use, and supported real-time, ready mission execution. The main weakness at this stage is the mission success rate, which remains below target because of technical issues associated with a still innovative and prototypal vehicle platform. For this reason, the current AC results should be interpreted as technically promising and operationally meaningful, while also clearly indicating that the next improvement cycle must prioritise vehicle stabilisation and deployment robustness before the replication pilot in Valencia.

4.2. Trustworthiness Assessment

4.2.1 Validity & Reliability

ATR

For ATR, Validity & Reliability is a core trustworthiness attribute because the practical value of the system depends on whether sonar-based detections remain technically correct, operationally stable, and dependable under realistic underwater conditions. In this use case, validation is not limited to nominal benchmark performance, but requires evidence that the full recognition chain (including data acquisition, annotation quality, model behaviour, runtime execution, and human supervision) supports trustworthy use in the pilot context.

The technical evidence collected during the pilot shows that the ATR system already reaches a strong level of validation on unseen pilot data. The high-frequency sonar model, evaluated on images never used during training or validation, achieved 95.6% precision, 97.6% recall, 96.0% F1-score, and 98.2% mAP@0.5, while maintaining real-time onboard performance at approximately 8 Hz on NVIDIA AGX Orin, against a sonar acquisition rate of 3 Hz. This is important because it shows that validation was carried out on genuinely unseen data and under realistic embedded-computing constraints. In addition, sonar remained operationally effective where optical sensing was severely limited, with optical visibility around 0.5 m and high-frequency sonar plus ATR remaining usable up to approximately 5 m. Validation was then extended through a dedicated robustness check on data from the last campaign, where targets had been repositioned differently in the water. This additional test is particularly relevant for Validity & Reliability because it probes the sensitivity of the model to changed scene geometry. In this condition, the model remained usable, but showed a more visible increase in false positives, while recall remained comparatively stable. This indicates that the current ATR branch is already reliable in-domain, while its main residual weakness lies in generalisation under changed pose and scene configuration.

The FAITH tools helped make this assessment more rigorous. The AI Data Passport documented the high-frequency FLS dataset as a governed asset by making explicit its provenance, annotation type, training role, access conditions, intended use and operational limitations. This is important for Validity & Reliability because it clarifies that the dataset supports close-range sonar-based instance segmentation in port environments and is not intended for autonomous critical decisions without human oversight or for use outside validated sensing conditions without further testing. The AI Model Passport did the same at model level, documenting the model identity, linked dataset, task, operational scope and main risk boundaries, including false positives, false negatives, reduced robustness outside the training domain, and operator overreliance. Together, the two passports strengthened the reliability assessment by defining

more clearly the operational envelope within which the model can be considered dependable. A further contribution came from TrustGuard, which shifted the perspective from raw performance values to asset-level trustworthiness reasoning. For the dataset, the tool highlighted that the central issue is whether the data are sufficiently controlled, representative and robust to justify reliable downstream model behaviour. For the model, TrustGuard identified the main reliability pressures as failure or malfunction of the ML application, changes in external conditions, and weaknesses in safeguards and software-component governance. This is important because it turns reliability into a managed lifecycle property rather than a one-off benchmark result.

Finally, trustSENSE, applied to the UNIFI AI Technical Team, supports the credibility of the current ATR validation process from the human side. The team reached a High trust level (3.80), with strong results in Technical Proficiency, Problem Solving, Collaboration, Proactivity and Threat Awareness, Responsibility and Ethics, and Motivation and Commitment. At the same time, more moderate scores in Resilience, Policy Adherence, Resource Accessibility, Integrity, Innovation and Adaptability, and Privacy and Compliance indicate the organisational areas that should be reinforced to maintain reliability over time. This is relevant because ATR reliability depends not only on model quality, but also on disciplined engineering, monitoring, retraining and controlled release practices.

AC

For AC branch, Validity & Reliability must be interpreted at system level. Indeed, the key question is whether the AC chain can consistently generate, update and execute survey trajectories while producing bathymetric outputs that are operationally useful and repeatable under real port conditions. In this case, validity is mainly supported by evidence on coverage completeness, mission execution, replanning responsiveness, and quality of the generated bathymetric outputs. From the operational-validation perspective, the current pilot already provides strong evidence of AC validity. The algorithm was tested on 20 distinct validation areas in the port environment, covering about 26,048 m² overall. Across all tested areas, the planner achieved the requested target coverage level in 100% of the cases. This is the strongest current indicator of AC validity, because it shows that the system can autonomously generate and execute routes that achieve the intended survey coverage under real pilot conditions.

A second important dimension is reliability of execution over time. During missions, the planner behaved in a real-time-compatible way: the next planning segment was computed before the vehicle reached the next waypoint, so the updated route was already available when needed. This means that replanning did not introduce perceptible waiting times during operation. At the same time, mission success remains a current limitation: 16 missions out of 20 were completed successfully, while 4 missions showed technical problems mainly related to the still prototypal

nature of the vehicle platform. This indicates that the current weakness lies less in the conceptual validity of the planning logic than in the reliability of the full integrated robotic chain. The quality of the bathymetric data further supports the validity side of the assessment. The AC system uses a Norbit WideBand Multibeam Sonar with 256–512 beams, 400 kHz nominal frequency with agility up to 700 kHz, adaptive ping rate up to 60 Hz, and depth resolution below 10 mm. In the current pilot, this configuration, combined with a vehicle speed of about 0.5 m/s, proved sufficient to generate dense bathymetric outputs and detailed seabed reconstructions suitable for coverage monitoring and practical survey use. Although a single formal geometric benchmark such as RMSE against a full external reference map was not extracted in this initial phase, the operational consistency and usability of the bathymetric products already support the validity of the AC function.

A complementary contribution comes from trustSENSE, again applied to the UNIFI AI Technical Team. Although trustSENSE does not assess the planner directly, it provides evidence on the maturity of the human asset responsible for AC development, debugging, deployment and refinement. The team reached an overall High trust level, with strong results in Technical Proficiency, Problem Solving, Collaboration, Proactivity and Threat Awareness, and Responsibility and Ethics. This is relevant because AC reliability depends heavily on iterative engineering work, including multibeam calibration, mission-log analysis, route-generation checks and refinement of the interaction between global and local planning. At the same time, the more moderate maturity dimensions (especially Resilience, Policy Adherence, and Resource Accessibility) suggest that the organisational side of reliability can still be strengthened through more formalised processes, stronger recovery routines, and more structured engineering governance.

4.2.2 Explainability & Interpretability

ATR

For ATR, Explainability & Interpretability is particularly important because the system is used to support operators in safety-relevant underwater inspection tasks, where a detection is useful only if it can be understood, checked, and placed in context. In this pilot, explainability is therefore not treated as a purely theoretical property of the CNN, but as an operational requirement of the whole decision-support chain: the operator must be able to see why a target has been detected, compare the evidence with known examples, and relate that detection to the mission situation in which it occurred.

This requirement has been translated into concrete pilot functionality. The ATR branch relies on Grad-CAM heatmaps to show the image regions that most influenced the prediction, and on retrieval of similar same-label training examples to support side-by-side comparison with known patterns. In practice, the operator does not receive only a target label, but also a structured

explanation package including the model name, target class, vehicle position, depth, yaw, original and detected images, Grad-CAM visualizations, and retrieved examples from the training dataset. The same workflow supports automatic PDF report generation, which means that explainability is not limited to a transient runtime visual aid, but becomes a reusable artefact for post-mission review, inspection reporting, and traceability. This is especially relevant in the underwater robotics context because the operator often has to interpret detections under uncertain sensory conditions. Grad-CAM improves explainability by highlighting what drove the model output, while the retrieved examples improve interpretability by helping the operator judge whether the current observation resembles known target patterns from the training set. The structured report adds a third layer by embedding the detection in its mission context, so that the AI output can be understood not as an isolated event but as part of a broader inspection sequence. In this sense, explainability in ATR is not reduced to "opening the black box," but becomes part of the practical interaction between the AI, the operator, and the mission.

The FAITH tools have had a direct influence on the development of this attribute. The strongest contribution came from TrustGuard, which explicitly identified Explainability & Interpretability as one of the most exposed trustworthiness dimensions for the ATR model. In the in-isolation assessment, the threat related to insufficient safeguards rises to a very high level under this attribute, making it clear that explanation quality is not a cosmetic improvement but a core requirement for trustworthy deployment. This changed the role of explainability in the pilot: instead of remaining an optional enhancement, it became a concrete mitigation priority. The use of Grad-CAM, similar-example retrieval, and richer operator-facing outputs should therefore be read not only as interface improvements, but as direct responses to the trustworthiness risks highlighted by the tool.

The Data Passport and AI Model Passport contribute in a more indirect but still important way. They improve interpretability at lifecycle level by documenting what the model is, which dataset supports it, what its intended use is, and where its operational limits lie. This means that interpretability is strengthened not only at prediction level, through visual explanations, but also at governance level, because the model and its data are easier to understand as bounded and documented assets. In particular, the fact that the model is explicitly linked to a specific training dataset and to a defined use scenario helps avoid overinterpreting its outputs outside the validated scope.

A further contribution comes from trustSENSE, which supports this dimension through the maturity of the human team. Explainability mechanisms are only useful if the technical team is able to design them well, interpret them correctly, and refine them in response to user feedback. The high trust level obtained by the UNIFI AI Technical Team therefore supports confidence that the current explanatory workflow is not only technically implemented, but can also be maintained

and improved over time. At the same time, the more moderate dimensions around resilience, policy adherence and resource accessibility suggest that these explanatory processes can still become more systematic in future iterations, especially in terms of standardisation and formal reporting.

AC

Here the issue is not explaining a predicted class, but making the behaviour of the planner understandable to the operator: why a route was selected, how the coverage area is being traversed, how risk areas and obstacles affect the mission, and how the final bathymetric result should be interpreted. In this sense, explainability in AC is mainly linked to route transparency, mission-state visibility, and traceability of planning decisions.

The current AC implementation already provides a reasonably strong explanatory layer for the present pilot stage. The operator can define the mission polygon directly on the map, insert risk areas and obstacles, visualise the generated path, follow coverage progression in real time, and observe the evolution of bathymetric outputs during execution. This is important because it means that the planner is not operating as an opaque autonomous block. Instead, its behaviour is exposed through a visible mission-management environment in which the operator can see what the system is doing and why. The explanatory value here comes from the combination of map-based interaction, real-time route visualisation, and continuous mission feedback. A key aspect is that interpretability begins already before route execution. The operator does not simply launch a generic autonomous mission; the survey area, requested coverage, risk zones and obstacles are all explicitly defined and displayed. As a result, the generated route can be interpreted in relation to inputs that are visible, editable and operationally meaningful. This is especially relevant in a port context, where route choices must be understandable not only to developers but also to users who are responsible for mission supervision and, if necessary, intervention.

trustSENSE adds the human-side perspective here as well. The maturity of the UNIFI AI Technical Team is relevant because the explanatory workflow in AC depends on engineering choices: how clearly the GUI presents route updates, how well logs are structured, and how easily operators can reconstruct what happened during the mission. The high trust level of the team therefore supports confidence that this dimension can be sustained and improved, although more formalisation will still be useful as the system moves toward broader operational use.

4.2.3 Safety

ATR

For ATR, Safety concerns the extent to which the system avoids contributing to unsafe inspection outcomes, such as missed hazardous objects, unnecessary interruption, or unsafe operator overreliance on model outputs. In this pilot, safety is not defined only by nominal detector accuracy, but by the combination of model performance, runtime safeguards, explicit operating boundaries, and preserved human supervision. The first layer of safety evidence comes from technical performance on pilot data. On unseen data from the first three campaigns, the high-frequency ATR model achieved 95.6% precision, 97.6% recall, 96.0% F1-score and 98.2% mAP@0.5, with relatively low false positive and false negative rates under nominal conditions. This is relevant for safety because it shows that, within the validated pilot domain, the system can limit both false alarms and missed detections to a reasonably low level. At the same time, the additional robustness-oriented evaluation on the last campaign showed that, when targets were repositioned, the system remained usable but became more permissive, especially through a more visible increase in false positives. This is an important safety finding, because it shows that the pilot does not rely only on nominal metrics, but also acknowledges degraded performance when scene conditions move away from the original data distribution.

A second and more explicitly safety-oriented element is the introduction of an OOD-aware runtime safeguard. The current ATR branch includes an out-of-distribution detection mechanism that acts as a trustworthiness add-on to the YOLO-based sonar pipeline. In practical terms, live detections are compared against the training distribution so that unusual or poorly represented inputs can trigger graded responses rather than being treated as ordinary detections. This is crucial from a safety point of view because it directly addresses one of the most important hazards identified in the trustworthiness analysis: the system should not behave as if it were fully confident when it is operating outside the data conditions on which it was validated. This safety logic is reinforced by the evolved application architecture. The current ATR software includes alarm states, acknowledgement procedures, pause mechanisms, and operator intervention logic, meaning that safety is not left to the model output alone. The GUI exposes mission and target information in a way that allows the operator to inspect the event in context and intervene when required. In this sense, the ATR branch is now much closer to a supervised decision-support system than to a standalone detector. That is particularly important in the port domain, where a safe AI system is one that can be interrupted or overruled in time, not simply one that produces accurate scores.

The FAITH tools made this safety posture more explicit and more disciplined. TrustGuard identified the main safety-relevant failure modes and helped prioritise countermeasures around malfunction, distribution shift, and inadequate safeguards. Data Passport and AI Model Passport

contributed by constraining the safe operating envelope: both make clear that the dataset and model are intended for validation and pilot use in bounded underwater conditions and are not intended for unrestricted autonomous critical decisions. This is directly relevant to safety because it prevents overclaiming what the current system can safely do.

Finally, trustSENSE adds the human-asset perspective. The high trust level of the UNIFI AI Technical Team supports confidence that the safety mechanisms can be maintained through retraining, threshold revision, anomaly analysis and controlled system updates. At the same time, the more moderate maturity dimensions around resilience and policy adherence indicate that the safety-management process can still become more formal and repeatable.

AC

For AC, Safety concerns the ability to execute bathymetric survey missions in port waters while minimising hazards for the vehicle, surrounding infrastructure and human operators. Here safety is not centred on recognition performance, but on whether the planning-and-execution chain can avoid unsafe routes, react to unexpected obstacles, remain inside acceptable operational boundaries, and preserve human authority whenever the situation becomes uncertain. The safety architecture adopted in the pilot is based on a two-tier planning structure. At the global level, the planner uses an RRT*-based strategy that takes into account both the mapping area and the risk zones defined by the operator, so as to maximise useful coverage while reducing exposure to hazardous regions. At the local level, an elastic-band planner adjusts the route between waypoints in real time and reacts to unexpected hazards such as unforeseen obstacles. This is a strong safety feature because it means that protection is embedded in the logic of route generation and replanning itself, rather than added only afterwards through manual supervision.

The pilot implementation further strengthens this through the operator interface. The operator can define the survey polygon, desired coverage level, risk areas and obstacles, and can monitor the route, coverage progression and bathymetric evolution during the mission. Most importantly, the operator can halt the mission or switch to teleoperated control at any time. This makes the human-on-the-loop logic central to the safety case: autonomy is used to improve efficiency and reduce exposure, but final authority remains with the operator. The communications architecture (Ethernet tether to a passive buoy, local Wi-Fi to quay operators, and 4G streaming to remote stakeholders) also contributes by ensuring continuous situational awareness and timely intervention capability.

Pilot evidence supports a positive reading of this attribute. No diver was involved in the routine execution of AC missions, and no operator was required onboard a support vessel during the survey itself. This is one of the clearest safety benefits of the AC branch: it reduces direct human exposure to hazardous underwater conditions compared with more traditional survey

procedures. At the same time, safety cannot yet be considered fully mature, because mission success still depends on the stability of a vehicle platform that remains partly prototypal. This means that the planning logic is already safety-oriented, but the full system still requires further hardening before broader deployment.

trustSENSE complements this by showing that the technical team has the maturity needed to sustain the safety case, particularly in multibeam calibration, risk-zone handling, route interpretation and system refinement. However, as with ATR, the more moderate maturity dimensions suggest that the safety-management process can still be made more formal and repeatable.

4.3 Preliminary Impact Evaluation

The preliminary impact of the AI systems in the underwater robotics LSP can already be assessed qualitatively on the basis of the evidence collected during the initial pilot, including stakeholder engagement, the workshop in Ravenna, field preparation and deployment activities, and the first operational adaptations introduced during execution. At this stage, the assessment remains necessarily preliminary, since the full replication cycle has not yet been completed and some scenarios have not yet been validated in real operating conditions. Nevertheless, the current evidence already supports a meaningful evaluation of both the operational and technical impact of the ATR and AC systems, including both positive and limiting aspects, as well as actual and still-potential effects.

Operational impact

From an operational perspective, the most visible positive impact of the pilot is the progressive demonstration that AI-enabled AUVs can support port inspection and monitoring activities in a way that is more systematic, less intrusive, and potentially safer than traditional diver-based or vessel-based procedures. Even before routine deployment, the pilot has already helped frame underwater inspection and survey operations not only as isolated engineering tasks, but as structured decision-support services embedded in port safety and infrastructure-management workflows.

A first concrete operational impact concerns stakeholder awareness and readiness. The Ravenna workshop created a shared understanding among the Port Authority, end users, researchers and other actors of what ATR and AC can realistically do in port conditions, which trustworthiness dimensions matter most, and how human supervision is preserved in practice. The technical sessions and interactive discussions on explainability, robustness and safety contributed directly to more informed expectations among users. This is important because trust and adoption in port

environments depend heavily on whether stakeholders understand how the systems behave, what outputs they produce, and how intervention remains possible at all times.

A second impact is the emergence of a human-on-the-loop operational model that appears well aligned with stakeholder expectations. The pilot activities, including field preparation and training, reinforced this architecture as a practical compromise between automation and operational control. Rather than replacing users, the systems support them in supervising, interpreting and reacting more efficiently to underwater events. This is particularly relevant in port operations, where resistance to adoption is often linked to concerns about full autonomy. Presenting the systems as operator-support tools, rather than autonomous replacements for human judgement, has therefore had a positive effect on acceptance and perceived usefulness.

A third positive impact, still preliminary but already visible, concerns the potential reduction of risky human exposure. The pilot has shown a credible pathway toward diverless, or at least diver-reduced, operations in selected tasks, especially for wide-area reconnaissance and early screening. This is particularly evident for ATR, where the vehicle can inspect an area with sonar before a closer approach is made, and for AC, where the system can map a selected region while the operator supervises from shore. The operational significance of this impact is high, even if it is not yet fully quantified.

At the same time, some limiting operational impacts also emerge. The most important is that the real deployment environment does not always provide naturally occurring conditions that match all intended scenarios directly. As a result, the pilot has had to rely in part on controlled or adapted scenarios, such as synthetic underwater targets, in order to preserve the credibility of the tests. While this does not weaken the value of the pilot, it does limit the extent to which the current results can already be interpreted as a full demonstration of all intended use cases in routine port practice. A second cautionary point is the continued dependence on logistics, authorisations, environmental conditions and coordination with port stakeholders. The pilot confirms that underwater AI deployment in port areas is not only a technical issue, but also a matter of feasibility within a complex operational environment. Even technically mature systems may therefore face non-negligible barriers to routine use when port traffic, testing permissions or seabed conditions limit the possibility of performing representative missions.

Technical impact

From a technical perspective, the pilot is already producing substantial impact by demonstrating the feasibility of integrating AI models, sensing systems, operator interfaces and AUVs into a coherent underwater inspection and surveying pipeline. The current results show that the intended technical direction of both ATR and AC is well founded and is beginning to materialise in meaningful ways.

For ATR, one of the most relevant positive impacts is the consolidation of a practical sonar-based recognition pipeline for underwater objects of interest. The system combines CNN-based recognition, target geo-localisation, risk-oriented categorisation, and operator feedback through the GUI. The pilot has further refined this by introducing a dual-inspection logic, in which a lower-frequency sonar supports long-range search and a higher-frequency sonar supports short-range, higher-resolution confirmation. This is an important technical refinement because it aligns sensing strategy with real operational needs: broad search and local verification are treated as two connected but distinct perception stages. A second positive impact for ATR concerns the integration of trustworthiness-oriented design choices into the technical system. The pilot has translated stakeholder concerns into concrete measures such as out-of-distribution detection, uncertainty-aware support, real-time visualisation of detections, heatmap-based explanation of model outputs, and comparison with similar training examples. This is technically important because it means that the system is being improved not only in terms of raw performance, but also in terms of robustness and comprehensibility. In a safety-relevant port setting, these properties are essential for practical usability.

For AC, the technical impact is already visible in the progressive validation of a coverage path-planning framework for autonomous bathymetric mapping. The planning strategy is framed not only as an efficiency mechanism, but also as a safety-aware approach combining a global risk-aware planner with local adaptation for unexpected obstacles. This is a significant contribution because it connects path efficiency, mapping completeness and safety within a single operational framework, moving the system beyond simple autonomous motion toward context-sensitive mission planning for port surveys. Another positive technical impact concerns the increasing maturity of the operator interfaces and real-time visualisation layer for both systems. The pilot has shown that technical performance alone is not sufficient: usable, interpretable and responsive interfaces are part of the technical system itself. In this sense, the pilot has validated not only the core AI algorithms, but also the wider chain through which those algorithms become actionable for end users.

There are, however, also limiting technical impacts that must be acknowledged. For ATR, the main limitation remains the scarcity of representative underwater datasets and the difficulty of validating the system on naturally occurring anomalies in real port conditions. While synthetic targets and controlled deployments are technically valuable, they cannot fully replace repeated exposure to diverse real-world anomalies. This limits the current generalisability of the system and means that further field data collection and fine-tuning remain necessary. For AC, the main technical limitation is that the change-detection functionality has not yet been meaningfully validated, because the pilot environment did not provide sufficient bathymetric change between surveys. As a result, the technical evidence currently demonstrated for AC mainly concerns

coverage planning and bathymetry generation, rather than repeated-survey comparison and anomaly highlighting over time. This does not reduce the value of the core system, but it does mean that the full intended capability is only partially evidenced at this stage.

A further challenge for both systems is the continued dependence on environmental variability and system-integration constraints. Underwater sensing remains affected by turbidity, low visibility, acoustic noise, and the practical limits of onboard computing and communications. The pilot has confirmed that robustness depends not only on model quality, but also on calibration, communications stability, real-time processing constraints, and mission-specific adaptation. A substantial part of the technical impact of the pilot therefore lies in exposing and progressively reducing these integration bottlenecks.

Overall preliminary assessment

Taken together, the current evidence suggests that the AI systems are already generating a positive preliminary impact in the LSP scenario, both operationally and technically. Operationally, they are strengthening stakeholder engagement, enabling a realistic human-on-the-loop model for underwater inspection and survey operations, and opening a credible path toward safer and less disruptive workflows. Technically, they are validating the integration of underwater AI perception and autonomous planning with real vehicles, sensors and operator interfaces, while also showing that trustworthiness requirements can be translated into concrete design choices. At the same time, the assessment also highlights the current limitations of the initial pilot, especially the incomplete coverage of all intended scenarios, the dependence on environmental and logistical conditions, and the need for further data collection and field validation. For these reasons, the present impact should be interpreted as promising but still evolving, with the replication pilot expected to play a decisive role in confirming, extending and generalising the benefits observed so far.

5. Regulatory and Ethical Assessment

5.1. Regulatory Alignment with the EU AI Act and Other Relevant Frameworks

The Initial Pilot in the Port of Ravenna is designed to align with the EU Artificial Intelligence Act (AI Act) and other relevant regulatory frameworks, ensuring that the AI-driven underwater robotic solutions deployed meet the highest standards of trustworthiness, transparency, and compliance. As a high-risk AI system in the maritime domain, the pilot must comply with the stringent requirements outlined in the AI Act, particularly those related to risk management, data governance, and human oversight.

Current Alignment Status

The FAITH project, through its consortium partners, has conducted a comprehensive gap analysis to assess the alignment of AI systems with the EU AI Act. This analysis confirms that the SUT comply with the foundational principles of the AI Act, including:

- *Risk Management*: The FAITH AI Trustworthiness Assessment Framework (AI TAF) has been adapted to incorporate the AI Act's risk management requirements. The AI TAF aligns with the NIST AI Risk Management Framework (RMF) and ENISA's guidelines, providing a structured approach to assess and manage risks throughout the AI lifecycle.
- *Data Governance*: The AI models used in the pilot adhere to the EU AI Act's requirements for high-quality data inputs and transparency, aligning with the GDPR and other data protection regulations. The AI Model Passport within the FAITH AI TAF ensures traceability and documentation of data flows and model decisions.

The alignment process also considers the HLEG Guidelines on Trustworthy AI, which emphasize principles such as human agency, technical robustness, privacy, transparency, fairness, societal well, being, and accountability.

High-risk AI System Requirements

The EU AI Act classifies AI systems used for critical infrastructure, such as port operations, as high-risk. To meet these requirements, the pilot in Ravenna has implemented the following measures:

- Risk Assessment and Mitigation: The FAITH AI TAF is used to conduct thorough risk assessments, identifying vulnerabilities and implementing countermeasures. This process aligns with the AI Act's requirement for continuous monitoring and evaluation of AI systems.
- Transparency and Explainability: The AI models deployed in the pilot provide clear explanations for their decisions, ensuring compliance with the AI Act's transparency requirements. This is achieved through interpretable AI techniques and the AI Model Passport, which documents the rationale behind model decisions.
- Accountability and Governance: The pilot includes robust governance mechanisms to ensure accountability for AI-driven decisions. This includes the establishment of an AI oversight committee and the implementation of audit trails for all AI, related activities, in line with the AI Act's emphasis on accountability and transparency.

FAITH's Support in Regulatory Alignment

The FAITH project provides a structured approach to regulatory alignment through its AI Trustworthiness Assessment Framework (AI TAF) and associated tools. These resources support the FAITH partners responsible for the underwater robotics pilot (WP7) in:

- Conducting Regular Audits: The AI TAF enables partners to perform regular audits of the AI systems, ensuring ongoing compliance with the EU AI Act and other frameworks. The framework includes predefined metrics and controls to evaluate the trustworthiness of AI systems against the AI Act's requirements.
- Documenting Compliance: The AI Model Passport and other FAITH tools facilitate the documentation of compliance efforts, providing a clear record of the steps taken to meet regulatory requirements.
- Engaging with Regulators: FAITH supports partners in engaging with regulatory authorities, such as the European Commission and national competent bodies, to ensure that the pilot's alignment efforts are recognized and validated.

Risks and Mitigation Strategies

While the pilot is progressing well in terms of regulatory alignment, the following risks have been identified:

- Regulatory Updates: The EU AI Act is a dynamic framework, with potential updates or clarifications that could impact the pilot's compliance status. The FAITH consortium is actively monitoring these developments and participating in relevant consultations.
- Operational Challenges: Ensuring compliance with the AI Act's requirements for high-risk AI systems may introduce operational challenges, particularly in areas such as data quality and human oversight. The pilot includes contingency plans to address these challenges.

5.2. Port-specific and Local Compliance

The initial pilot was carried out in full consideration of the operational, regulatory and logistical constraints of the Port of Ravenna. In practical terms, port-specific and local compliance was not treated as a separate administrative formality, but as a condition embedded in site selection, mission planning, stakeholder coordination and field execution. The Ravenna pilot was organised only in port areas where the activities could be performed safely, with limited or no interference with ordinary navigation and port operations, and with the required level of local authorisation depending on the selected site.

At legal and procedural level, the pilot operated in accordance with authorisations issued by the Port Authority and by the Maritime Authority, and followed the Temporary Operational Guidelines applicable to underwater inspections. This is the main local compliance framework for the pilot, because it ensures that underwater robotic operations are conducted under the same port, governance logic that applies to other safety-relevant activities in the harbour domain. In addition, the local compliance assessment confirmed that the pilot takes place within port waters and therefore remains aligned with the specific environmental and maritime, regulatory context of the port area, while also complying with local port environmental policies.

The site, selection strategy in Ravenna further demonstrates this port-specific compliance approach. The city dock was identified as a low-impact area where navigation is forbidden, making it suitable for test activities without interference with normal port operations; in that site, no additional authorisation was needed and Port Authority premises were available for electricity, data connection and storage of devices. By contrast, the fishing boats and patrol boat dock and the southern breakwater were both associated with Harbour Master authorisation requirements, because they are closer to active navigation contexts and therefore require a stricter operational, control framework. In both cases, the pilot was planned so as to maintain low impact on navigation and preserve safety conditions during deployment.

Operational constraints specific to the port environment were also explicitly taken into account. The selected Ravenna sites are characterised by low or very low underwater visibility and by different degrees of interaction with surrounding maritime activity. These conditions influenced not only the technical setup of the pilot, but also the compliance logic of field operations, since they required mission planning that remained compatible with safety expectations, local traffic conditions and the physical characteristics of the deployment areas. The availability of port, side infrastructure such as electricity, data connectivity, storage space and operational support was also a relevant local condition, because it enabled repeated deployments while keeping the pilot inside a controlled and manageable operational setting.

Another important aspect of local compliance concerns the role of port stakeholders in supervising and validating the appropriateness of the activities. The Port Authority's infrastructure, maintenance and safety/security functions, together with Harbour Master / Maritime Authority / Port Control, are identified as relevant actors in the pilot context. Their involvement is important not only from an organisational perspective, but also from a compliance perspective, because it ensures that the pilot remains aligned with local safety expectations, port, operational priorities, and the requirement that regulatory bodies be informed about activities underway in the port. This stakeholder structure strengthens the local legitimacy of the pilot and reduces the risk of a gap between technical testing and port, governance requirements.

Overall, the initial pilot shows a good level of port-specific and local compliance. The activities were conducted under the relevant port and maritime authorisation framework, were adapted to the operational constraints of the selected Ravenna sites, were planned to avoid undue interference with navigation and routine port functions, and were supported by direct involvement of the competent port actors. In this sense, compliance at local, port level was achieved not only through formal permissions, but also through a broader operational discipline that integrated safety, site suitability, stakeholder coordination and respect for the practical constraints of the harbour environment.

5.3. Technical Compliance of Data and AI Models

From a technical-compliance perspective, the initial pilot shows that the ATR data and model assets have been managed through a structured combination of dataset governance, model traceability, documented intended use, and trustworthiness-oriented technical assessment. In the current pilot, compliance is not based on a standalone certification step, but on the coordinated use of FAITH tools to document provenance, define operational boundaries, support reproducibility, and assess technical risks affecting the AI lifecycle. This is particularly relevant in the underwater robotics context, where reliability, safety, transparency and accountability depend not only on model performance, but also on how data and models are governed over time.

For the data asset, UNIFI used the Data Passport in the FAITH AI Model Hub to document the high-frequency FLS dataset used in the ATR pipeline. The passport provides a structured digital identity for the dataset by recording its provenance, ownership, lifecycle role, intended and non-intended uses, access conditions, and security/governance measures. In practical terms, this means that the dataset is not treated as an informal collection of annotated sonar images, but as a governed asset linked to a specific supervised-learning workflow and to a clearly delimited operational context. This contributes to technical compliance because it makes explicit under which conditions the dataset can be used, who is responsible for it, and what restrictions apply to its reuse.

For the model asset, UNIFI used the AI Model Passport to document the high-frequency ATR model associated with the study FAITH WP7 FLS Instance Segmentation Study. The passport records the model identity, technical characteristics, linked training dataset, task definition, intended use, and main trustworthiness limits. This is particularly important because it creates a documented connection between the training data and the resulting model, thereby strengthening lineage, reproducibility and accountability across the ATR development chain. The model is explicitly framed as a decision-support component for research, validation and pilot evaluation in port underwater environments, rather than as an unrestricted autonomous system.

From the perspective of responsible data and model management, an important compliance feature of the current pilot is the presence of explicit operational limits. The dataset is documented as suitable for training and validation of sonar-based target instance-segmentation models in bounded port scenarios, under restricted access and with defined organisational, legal and technical safeguards. Likewise, the model documentation makes clear that the main technical concerns include false positives, false negatives, reduced robustness outside the training domain, and operator overreliance. This is significant because technical compliance here is not reduced to secure storage or metadata completeness; it also includes a clear delimitation of what the data and model are meant to support, and under which conditions their use remains justified.

A further compliance layer is provided by the TrustGuard assessments carried out on both the HF FLS dataset and the trained ATR model. These assessments complement the passport-based documentation by examining the assets through structured questions on threats, consequences and vulnerabilities. For the dataset, this includes issues such as provenance, representativeness, data quality, and governance of the data pipeline. For the model, it includes aspects such as component integrity, monitoring, traceability, resilience to changing deployment conditions, and documentation quality. In this sense, the passports support the documentation and traceability layer, while TrustGuard supports the technical risk and control-assessment layer.

6. Conclusions and Roadmap for the Final Pilot Phase

6.1. Summary of achievements in the second period

During the second reporting period, the underwater robotics pilot achieved a substantial consolidation of both the technical deployment and the trustworthiness, management dimensions of the FAITH approach. The initial pilot in Ravenna did not remain at prototype, demonstration level, but evolved into a structured validation activity for the two AI SUT, ATR and AC, while also creating the methodological and operational basis for the second implementation cycle and the future replication pilot in Valencia. In this respect, the period can be considered successful both as a pilot, execution phase in its own right and as a preparatory phase for transfer, scaling and refinement.

A first major achievement concerns the progressive maturation of the AI systems. ATR evolved from an initial underwater, detection prototype into a more mature operator-centred, trustworthiness-oriented decision-support system, with stronger interface integration, richer contextual support for the operator, and more structured handling of detections and alarms. In parallel, AC progressed from a planning capability into an operational autonomous, surveying chain able to generate and execute coverage plans for real port areas and to support bathymetric mapping in field conditions. Taken together, these developments mean that the pilot was able to

move beyond conceptual feasibility and generate concrete evidence on the use of AI-enabled underwater robotics in realistic port workflows.

A second major achievement lies in the collection of meaningful technical evidence on pilot data. For ATR, the pilot succeeded in creating, annotating and exploiting dedicated sonar datasets and in training and evaluating both high-frequency and low-frequency FLS models. The resulting quantitative evidence showed strong performance on unseen pilot data, near-real-time onboard execution, and clear operational usefulness in visibility conditions where optical sensing alone was severely limited. For AC, field validation on multiple survey areas showed that the system could consistently achieve the intended coverage objective, generate useful bathymetric outputs, and support real-time, capable mission execution.

A third achievement concerns the human-centred and operational validation dimension of the pilot. The initial pilot did not focus only on technical deployment, but also on the engagement of relevant stakeholders and end users, the collection of usability and trustworthiness, related feedback, and the verification of whether the systems could be meaningfully inserted into real port workflows. Through workshops, field interaction, training and progressive involvement of operational actors, the pilot generated evidence not only about what the AI systems can technically do, but also about how they are perceived, supervised and used by the port community. This is essential in the FAITH context, because the core objective is not only to deploy AI in a critical domain, but to do so in a way that fosters practical trust and acceptance.

A fourth major achievement is the concrete application of FAITH trustworthiness tools to the underwater robotics pilot. During the reporting period, the pilot moved from general methodological alignment to actual tool use. The FAITH AI Trustworthiness Assessment Framework v1 was applied through in-isolation assessment of the main AI assets; TrustGuard was used to evaluate ATR, related data and model assets; trustSENSE was used to assess the maturity of the UNIFI AI Technical Team; and the FAITH AI Model Hub functionalities were used to register and describe the relevant ATR assets through the Data Passport and AI Model Passport mechanisms. This means that trustworthiness assessment in the pilot was not left at the level of abstract principles, but translated into a structured operational workflow producing concrete threat, risk and control evidence.

Closely related to this, another achievement is the establishment of a first technical, compliance and governance baseline for the ATR branch. Through the use of the Data Passport and AI Model Passport functionalities, the relevant assets were documented with explicit intended use, access conditions, provenance and operational limitations, while TrustGuard complemented this layer with structured assessment of threats, consequences, vulnerabilities and countermeasures. The overall result is that the initial pilot already provides a credible basis for data governance,

traceability, intended, use delimitation and broader trustworthiness management, even if further work is still needed to strengthen robustness under changing deployment conditions and to formalise some lifecycle controls more completely.

A further achievement is that the pilot has clarified the main residual gaps and next, step priorities with much greater precision than would have been possible through technical testing alone. The current evidence shows that the most relevant remaining issues concern validity and reliability, safety, accountability and transparency, and explainability and interpretability, especially in relation to data quality, robustness under changed environmental conditions, component governance and update-time assurance. This is itself a meaningful achievement, because the pilot has not only demonstrated strengths, but has also made the remaining weaknesses visible in a structured way, thereby allowing the next implementation cycle to focus on the most material risks rather than on generic refinement.

Finally, the second reporting period achieved a clear roadmap orientation toward the replication phase. The next phase is already framed as a validation and refinement step in a new operational environment, with preparatory work focused on system refinements, stakeholder engagement and regulatory approvals before the replication pilot in the Port of Valencia. In this sense, the initial pilot has already fulfilled one of its most important objectives: it has produced the technical, operational and trustworthiness evidence necessary to move from an initial deployment in Ravenna toward a more mature, comparative and scalable second, phase validation.

6.2 Plan for Replication Phase

The replication phase of the FAITH project's underwater robotics pilot is structured to validate and refine the AI-driven solutions deployed in the Initial Pilot (Port of Ravenna), incorporating lessons learned and iterative improvements. Based on the current status and planning outlined in Milestone 9 (LSP4), the following high-level timeline is proposed:

- **Q2 2026 – Q1 2027**

  Finalization of all preparatory activities, including system refinements, stakeholder engagement, and regulatory approvals. This phase ensures that the AI systems (ATR and AC) and operational workflows are optimized based on feedback from the Ravenna pilot and aligned with the FAITH AI TAF.

- **Mid-2027**

  Execution of the Replication Pilot in the Port of Valencia. This phase focuses on deploying the improved AI systems in a new operational

environment, validating their performance under real-world conditions, and assessing trustworthiness, reliability, and scalability. The pilot will include scenarios such as bathymetric analysis, seabed mapping, underwater infrastructure inspection, and vessel hull inspections.

- **Q4 2027 – Q1 2028**

  Evaluation and documentation of the replication pilot's outcomes, including a comparative analysis with the Ravenna pilot. This phase will generate comprehensive reports on trustworthiness improvements, operational lessons learned, and recommendations for future deployments.

Note: The timeline is indicative and subject to adjustments based on external factors such as final approvals from the Port Authority of Valencia and the acquisition of necessary operational authorizations.

6.3. Identified risks for the Replication Phase

The replication phase in the Port of Valencia is intended to validate and refine the ATR and AC solutions in a new operational environment, after the preparatory activities, system updates, stakeholder engagement and regulatory actions carried out following the Ravenna pilot. In this context, the main risks are not only technical, but also organisational, logistical and regulatory, because the success of the replication pilot depends on the ability to transfer the systems and workflows to a different port setting while preserving performance, safety and trustworthiness.

A first risk concerns authorisations, local permissions and operational coordination in the Port of Valencia. The replication timeline is explicitly subject to the final approvals of the Port Authority of Valencia and to the acquisition of the necessary operational authorisations. Any delay in these processes may shift the field, testing calendar, reduce the available experimentation window, or constrain the set of scenarios that can be executed. This is particularly relevant because underwater deployments in port areas depend on access to suitable zones, safety clearances, compatibility with navigation and port activities, and the availability of local logistical support.

A second major risk concerns transferability of the AI systems to a new deployment environment. The initial pilot already showed that the most relevant residual issues are related to robustness under changing deployment conditions, data quality, model generalisation, component governance, and update-time assurance. For ATR, this means that sonar-based target recognition may behave differently if the Valencia environment presents different seabed textures, background clutter, target configurations, turbidity patterns or acquisition geometries. For AC, the main risk is that route generation and bathymetric coverage, although already validated in Ravenna, may be affected by different terrain layouts, currents, obstacles, or operational

constraints in the new port. In both cases, there is therefore a concrete risk that performance in Valencia may not fully match the results obtained in Ravenna unless the systems are further adapted and revalidated beforehand.

A third risk concerns system stability and full, chain operational reliability, especially for the AC branch. The initial pilot showed that the coverage logic is technically promising, but that mission success is still influenced by the maturity of the overall vehicle platform and by integration issues typical of an innovative prototypal system. If these stability issues are not sufficiently reduced before Valencia, they may affect the continuity of field trials, increase downtime during the replication phase, and make it more difficult to collect consistent comparative evidence. For ATR, a related risk concerns the reliability of the complete operational chain, including sensing, onboard processing, communications, interface behaviour and human supervision, rather than the model alone.

A fourth risk concerns data representativeness and the adequacy of further model refinement. The replication phase foresees additional data collection, fine, tuning and field trials in Valencia. This means that the quality, diversity and annotation consistency of the new pilot data will remain critical. If the additional data are not sufficiently representative of the operational conditions in Valencia, the retrained or updated ATR models may remain vulnerable to false positives, missed detections, or degradation under unfamiliar target poses and scene conditions. More generally, there is a risk that the evidence base used for model updates may still be narrower than the variability encountered in the replication environment.

A fifth risk concerns human readiness, training and operational uptake by end users. The initial pilot clearly showed that technical performance alone is not sufficient: meaningful deployment also depends on whether operators understand the system behaviour, trust the outputs, and are prepared to supervise and intervene when needed. In Valencia, insufficient user preparation or delayed stakeholder engagement could reduce the quality of field validation, limit the usefulness of feedback, and weaken the human-on-the-loop safeguards that are central to both ATR and AC. This is especially important because the replication phase is meant not only to repeat the Ravenna experience, but to validate scalability and operational trustworthiness in a new context.

A sixth risk concerns regulatory and compliance evolution during the final pilot phase. The pilot is progressing within a dynamic regulatory environment, and possible updates or clarifications in the EU AI Act or in related compliance expectations may introduce additional documentation, control or governance requirements. Even where no fundamental barrier is expected, these changes may still create practical challenges in areas such as data quality assurance, human oversight, traceability and formal compliance evidence, especially as the systems move toward a more mature deployment stage.

Overall, the identified risks for the replication phase can therefore be summarised as follows: delays in permissions and local authorisations; incomplete transferability of Ravenna results to Valencia conditions; residual platform and integration instability; insufficient representativeness of additional pilot data for robust model updating; limited user readiness or stakeholder engagement; and evolving regulatory or compliance requirements. These risks do not undermine the feasibility of the replication phase, but they confirm that the Valencia pilot should be approached as a controlled validation and refinement step, in which technical adaptation, operational preparation, and trustworthiness governance must continue in parallel.

7. References

[1] Allotta, Benedetto, et al. "A low, cost AUV for patrolling and monitoring." Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment 231.3 (2017): 740, 749.

[2] Vangi, Mirco, et al. "Design, development, and testing of an innovative autonomous underwater reconfigurable vehicle for versatile applications." IEEE Journal of Oceanic Engineering (2025).