Ref. Ares(2026)4470124 - 30/04/2026

D5.2

Report on the initial pilot - Transportation LSP

Related Work Package	**WP5 – Implementation of transportation LSP**
Related Task	Task 5.2 – The initial pilot
Lead Beneficiary	MERMEC
Contributing Beneficiaries	CNR
0Document version	v.1.0
Deliverable Type	R
Distribution level	SEN
Contractual Date of Delivery	30/04/2026
Actual Date of Delivery	30/04/2026

Authors	F. Martinelli (MERMEC), A. Cantini (MERMEC) L. Volpicella (MERMEC), F. Fabrizio (MERMEC), L. Fabrizio (MERMEC), S. Colantonio (CNR), G. Del Corso (CNR), S. Gravili (CNR), G.R. Leone (CNR), D. Moroni (CNR), O. Papini (CNR), M.A. Pascali (CNR)
Contributors	
Reviewers	E. Tsalapati (ATC), Cristina Guseth(FH), Teodora Musat (FH)

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
v.0.1	TOC	26/02/2026
v.0.2	First draft	14/04/2026
v.0.9	Draft for internal revision	24/04/2026
v.1.0	Version updated with internal reviewers' comments	29/04/2026

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This deliverable contains original unpublished work except where clearly indicated otherwise.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

The Large-Scale Pilot (LSP) on Transportation focuses on deploying, evaluating, and refining Artificial Intelligence (AI) solutions within a privacy-preserving smart camera platform.

The overarching objective is to give the public transport managers tools that support strategic planning, enabling evidence-based decisions on network design, timetable optimization, infrastructure investments, and improve passengers' comfort on board.

Due to severe bureaucratic delays associated with obtaining authorizations from the national railway operator, **the System Under Test (SUT) was strategically transitioned from trains to a bus-based public transport network**. This pivot proved highly advantageous, offering a flexible, accessible, and scalable environment for data acquisition and SUT experimentation.

The initial pilot evaluates two core analytical functionalities. **The Passenger Counting functionality successfully reached a Technology Readiness Level (TRL) of 6/7**, demonstrating a high real-world accuracy of 94.1% (a mean relative error of only 5.9%). Conversely, the Transit Residual Detection functionality—aimed at identifying unattended luggage and garbage—is currently at TRL 3/5. Out-of-the-box models like YOLO26n struggled significantly with garbage detection, achieving only a 4% correct labelling rate in real-world scenes. To overcome this, the CNR team engineered a custom, domain-specific dataset comprising 50,000 images. This dataset utilizes a highly controlled multi-view laboratory setup and will be used jointly with advanced generative AI (Inpaint and Juggernaut models) to create photorealistic synthetic environments for future model fine-tuning.

A major milestone of this period was the practical instantiation of the FAITH AI Trustworthiness Assessment Framework (TAF V1). Technical teams rigorously applied the FAITH toolset across the AI lifecycle:

- *TrustSense* evaluations confirmed a moderate-to-high trustworthiness maturity level among the AI developers and operators.
- *Data and AI Model Passports* were successfully implemented on the FAITH AI Model Hub, guaranteeing transparency and traceability regarding data provenance and model architecture.
- *TrustGuard* systematically identified and signalled threats. While the overall risk profile of the SUT was categorized as "Low", TrustGuard successfully isolated high-frequency operational threats specific to the transit domain, such as physical 'Evasion' (e.g., severe camera occlusions) and environmental variability.

The system strictly adheres to privacy-by-design principles, verified through a comprehensive Data Protection Impact Assessment (DPIA). By processing video streams near real-time at the edge and extracting only anonymized, aggregated numerical data, the SUT eliminates the need to store personal video data. Consequently, the system complies fully with the General Data Protection Regulation (GDPR) and is officially classified as a "low-risk" system under the European Union Artificial Intelligence Act (AI Act), while still maintaining essential human-in-the-loop oversight. Furthermore, stakeholder consultations with 29 domain experts confirmed a strong

industry willingness (60.7%) to test trustworthy AI assessment tools, highlighting the lack of transparency as the primary barrier to widespread adoption.

Looking ahead, **the Replication Pilot phase will leverage the existing bus infrastructure**, effectively eliminating the need for hardware installations, and enabling rapid scaling. The system will be deployed across a massive fleet of approximately 1,000 vehicles operating in 12 Italian provinces. Future efforts will focus on massive scale-up, extensive algorithmic fine-tuning for transit residual detection, and continuous trustworthiness monitoring to mitigate identified scaling risks, such as model drift, environmental variability, and potential automation bias among human operators.

Table of Contents

1 Introduction 10
1.1 Scope and objectives of the 1st pilot in the Transportation LSP 10
1.2 Structure of the document 11
1.3 Evolution of the AI System Under Test (SUT) from initial 11
2 Progress on the AI system deployment(s) and execution of the Pilots 14
2.1 Status of the 1st Pilot execution 14
2.1.1 Deployment Context & Domain Specifics 14
Passengers Counting 16
Transit Residual Detection 19
2.1.2 Stakeholders workshop 23
2.1.3 Users 25
2.1.4 Training 26
2.1.5 Evaluation Process 26
2.2 Status of the Replication Pilot preparation/initialization 28
2.3 Deviations from the original deployment plan 29
2.4 Data collection throughout the 1st Pilot execution 31
2.5 Integration into LSP workflows: Lessons learnt 33
2.5.1 Issues encountered and solutions 34
2.5.2 Lessons Learned for LSP Workflow Integration 34
3 Application of FAITH Tools 35
3.1 Application of the FAITH AI TAF V1 methodology 35
3.1.1 Cartography 35
3.1.2 Threat Analysis 38
3.1.3 Impact Assessment 44
3.1.4 Vulnerability Assessment 47
3.1.5 Risk Analysis 48
3.1.6 Countermeasures 48
3.2 Tool Validation Planning & Training Processes 49
3.2.1 Dry-run 49
3.2.2 Plan development 59
3.2.3 Tool Training and validation 61

3.3 FAITH Tool Validation Activities 62
3.3.1 TrustSense 62
3.3.2 Data & AI model Passport in the FAITH AI Model Hub 64
3.3.2.1 Passenger Counting - Data Passport 64
3.3.2.2 Transit Residual Detection – Data Passport 64
3.3.2.3 AI Model passports 65
3.3.3 TrustGuard 67
3.4 Threat Analysis 68
3.4.1 Passenger Counting – Dataset and AI model 68
3.4.2 Transit Residual Detection - Dataset 70
3.5 Initial Risk Assessment results 72
3.5.1 Passenger Counting – Dataset and AI model 72
3.5.2 Transit Residual Detection - Dataset 75
3.6 Countermeasures 77
3.6.1 Passenger Counting 77
3.6.2 Transit Residual Detection 78
3.7 Formal Tool Feedback 78
4 AI System Evaluation 80
4.1 Quantitative metrics on pilot data 80
4.2 Trustworthiness Assessment 81
4.2.1 Privacy 83
4.2.2 Security & Robustness 83
4.2.3 Validation & Reliability 84
4.2.4 Transparency & Accountability 85
4.2.5 Explainability & Interpretability 85
4.3 Preliminary Impact Evaluation 86
5 Regulatory and Ethical Assessment 87
5.1 Alignment with the EU AI Act 87
6 Conclusions and Roadmap for the Final Pilot Phase 90
6.1 Summary of achievements in the second period 90
6.2 Plan for Replication Phase 90
6.3 Identified risks for the Replication Phase 91

7 References 93
8 Annex I – Passenger Counting Assessment Report 94
9 Annex II -Transit Residual Detection Assessment Report 150

List of Abbreviations

Abbreviation	Explanation
AI	Artificial Intelligence
AI_TAF	Artificial Intelligence Trustworthiness Assessment Framework
AI Act	Artificial Intelligence Act
AIDA	Artificial Intelligence Distributed Analytics
CCTV	Closed-Circuit Television
CNR	National Research Council of Italy
DITECFER	Distretto per le Tecnologie Ferroviarie, l'Alta Velocità e la Sicurezza delle Reti (Tuscany Railways Stakeholders Consortium)
DPO	Data Protection Officer
DPIA	Data Protection Impact Assessment
EU	European Union
FAITH	Fostering Artificial Intelligence Trust for Humans
FORTH	Foundation for Research and Technology – Hellas
GDPR	General Data Protection Regulation
KPI	Key Performance Indicator
ISO	International Standardization Organization
ISTI	Institute of Information Science and Technologies
IT	Information Technology
LSP	Large Scale Pilot
LSP2	Large Scale Pilot 2 (Transportation)
MERMEC	Mermec Engineering
SUT	System Under Test
TRL	Technology Readiness Level
VMS	Video Management System
YOLO	You Look Only Once (state of the art obj detector)
YOLOv8	YOLO AI Model version 8
YOLO26n	YOLO 2026 AI Model nano size

List of Tables

Table 1 - Expected TRL of system functionalities. 14
Table 2 - Stakeholders consultations in 2025. 14
Table 3 - Data of the sample validation test. 19

Table 4 - Summary of the sample validation test. 19
Table 5 - Out-of-the box Yolo26n performances in recognition of garbage and luggage items.. 20
Table 6 –Yolo26n performance are very poor on garbage detection in real scene 21
Table 7 - List of the SUT assets. 36
Table 8 - AI Act classification of criticality levels. 37
Table 9 - Mapping between identified threats, associated impacts and key dimensions. 40
Table 10. Mapping between DPIA threats listed in Table 9 and the threats in FAITH Golden Table 41
Table 11 - Threat frequencies 44
Table 12 - Business consequences for poor trustworthiness characteristics..... 47
Table 13 - Threats and vulnerability of the AI assets related to the Transit Residual Detection (in the design phase) as evaluated during the Dry-run phase. 54
Table 14. Threats and vulnerability of the AI assets related to the Passenger Counting (in the develop and deploy phase) as evaluated during the Dry-run phase. 59
Table 15 - Training and support activities. 60
Table 16 - FAITH tools implementation and iteration 2026. 61
Table 17 – Prioritization of Trustworthiness attributes in the LSP on public transport. 82

List of Figures

Figure 1 - The system architecture in the edge-oriented configuration. 12
Figure 2 - The new experimental bus platform currently in use for the initial pilot 15
Figure 3 - The analytical functionalities are now executed at server level. 15
Figure 4 - The Passenger Counting functionality schema. 16
Figure 5 - Passenger counting on board. 17
Figure 6 - Passengers boarding time. 17
Figure 7 - The comparison of manual and AI count in the sample validation test. 19
Figure 8 - Performance issues of YOLO26n in the identification of garbage and luggage. 21
Figure 9 - Acquisition setup including rotating platform and adjustable artificial lighting. 21
Figure 10 - Example of multiple views (72) for each item (e.g., red luggage in this image). 22
Figure 11 - Multiview prototype ≈ 50000 images of 300 transit related items. 22
Figure 12 - Example of physically informed Generative image obtained by Inpaint pipeline. 23
Figure 13 - The targeted expert consultation carried out. 24
Figure 14 – Stakeholders' willingness to experiment with trustworthy AI solutions. 25
Figure 15 - The SUT results are compared with the ground truth observations. 27
Figure 16 - Example of total occlusion of the camera view. 27
Figure 17 - Example of partial malfunctions of the camera sensor. 28
Figure 18 - The future fleet will consist of about 1000 vehicles in 12 Italian provinces. 29
Figure 19 - Time schedule variation presented in Deliverable D5.1. 30
Figure 20 - New Time schedule variation. 30
Figure 21 - Images of the reference dataset are annotated using CVAT 33
Figure 22 - The FAITH TAF V1 methodology. 35

Figure 23 - TrustSense: samples of results collected after the use of the tool. 62
Figure 24 - TrustSense: results collected in a Domain AI User team (bus company in Italy). 63
Figure 25 - TrustSense: some examples of recommended mitigation actions. 63
Figure 26 - The Data Passport for the Passenger Counting on the FAITH AI Model Hub. 64
Figure 27 - The Data Passport for the Transit Residual Detection on the FAITH AI Model Hub. . 64
Figure 28 – AI Passport: execution of the "preprocess" stage of the pipeline. 65
Figure 29 - AI Passport: execution of the "train" stage. 66
Figure 30 - AI Passport: the training is completed. 66
Figure 31 – AI Passport: the model has been registered successfully on MLFlow. 67
Figure 32 - AI Passport: The LSP2_passenger_counter runs can be checked on MLflow. 67

1 Introduction

The Large-Scale Pilot (LSP) on public Transportation is focused on deploying, evaluating and further tailoring the Artificial Intelligence (AI) solutions embedded in a scalable privacy preserving platform of smart cameras for monitoring the efficiency of public transit and the safety and security of passengers on board of buses and trains. The work leverages the methods already developed by the involved partners (CNR and MERMEC Engineering) into an ad-hoc deployed scalable infrastructure in different pilot sites, supplied by the Italian National Railway Operator and Italian bus-line company. The work is organized in three steps:

Initially, the activities focused on the set-up of the smart sensor platform and the instantiation of the FAITH risk-based dynamic risk management approach for the specific transport domain, by defining the specific profile, potential vulnerabilities and potential misuses: this phase terminated one year ago in March 2025 and the report of the activities can be found in Deliverable D5.1.

Afterwards, the initial pilot phase refines and test the existing AI-powered methods, to identify and prioritize the specific trustworthiness requirements, and put in place measures against adversarial or noxious uses: this phase terminates in April 2026 and the report of such work is the focus of this document Deliverable D5.2.

Finally, the replication pilot implements the system on a larger scale to validate the refined trustworthiness characteristics, learnt from the initial piloting phase. This will be documented with Deliverable D5.3.

1.1 Scope and objectives of the 1st pilot in the Transportation LSP

The primary scope of the first pilot within the Transportation LSP is the deployment, evaluation, and refinement of AI solutions within a privacy-preserving smart camera platform. The main objective is to provide the public transport managers with tools that support strategic planning, enabling evidence-based decisions on network design, timetable optimization, infrastructure investments, and improve passengers' comfort on board.

Specifically, the System Under Test (SUT), focuses on the validation of two core analytical functionalities: real-time Passenger Counting and Transit Residual Detection, which encompasses the identification of unattended luggage and garbage.

The fundamental objectives of this initial pilot phase are threefold:

To test and refine the AI models in a real-world operational environment. Due to bureaucratic delays with the national railway operator, this environment was strategically redirected towards a bus fleet, thereby ensuring greater flexibility, scalability, and accessibility for data acquisition.

To apply the tools developed within the FAITH (such as TrustGuard, TrustSense, and the Data & AI Model Passports) to instantiate a dynamic risk management approach, assessing vulnerabilities, threats, and the overall trustworthiness level of both the technical team and the AI system. This will lead to a richer understanding of trustworthiness values and requirements

To gather direct feedback from the field through domain stakeholder consultations, establishing the groundwork for the large-scale integration anticipated in the subsequent "Replication Pilot" phase.

1.2 Structure of the document

This document is structured as follows:

Chapter 1 (Introduction): introduces the pilot, illustrating its scope and objectives, and tracing the evolution of the System Under Test (SUT) following the transition from the railway domain to the bus network.

Chapter 2 (Progress on the AI system deployment and execution of the Pilots): details the execution status of the first pilot, examining the deployment context, stakeholder and user involvement, data collection processes, deviations from the original deployment plan, and the lessons learned during integration into LSP workflows.

Chapter 3 (Application of FAITH Tools): provides an in-depth description of the practical application of the FAITH AI Trustworthiness Assessment Framework (TAF) V1 methodology. This section covers all phases of risk analysis (Cartography, Threat Analysis, Vulnerability and Impact Assessment, and Countermeasures) and documents the use of the FAITH AI Model Hub the validation activities conducted via TrustSense and TrustGuard tools.

Chapter 4 (AI System Evaluation): focuses on the evaluation of the system's quantitative metrics and trustworthiness requirements, analysing key attributes such as privacy, security, robustness, validation, transparency, and explainability, in addition to providing a preliminary impact evaluation.

Chapter 5 (Regulatory and Ethical Assessment): addresses the regulatory and ethical compliance of the system, discussing the Data Protection Impact Assessment (DPIA) and the degree of alignment with the requirements of the new EU AI Act.

Chapter 6 (Conclusions and Roadmap for the Final Pilot Phase): summarizes the achievements reached in this period, outlines the action plan for the "Replication Pilot" phase, and identifies potential future risks associated with it.

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

The core technology used in this pilot is AIDA platform (Artificial Intelligence Distributed Analytics) developed by MERMEC Engineering; AIDA is a scalable, privacy-preserving system that processes video streams collected from onboard Closed-Circuit Television (CCTV) cameras on public transport vehicles.

The SUT is composed of a selected subset of functionalities offered by the AIDA platform:

- *Passengers counting:* real-time counting of passengers on board

- *Unattended objects detection:* identifying unattended or lost objects in the passenger areas of the vehicles
- *Garbage identification*: detecting garbage, to assess the cleanliness levels of seats and carriages

The last two functionalities are merged into one task called *Transit Residual Detection*.

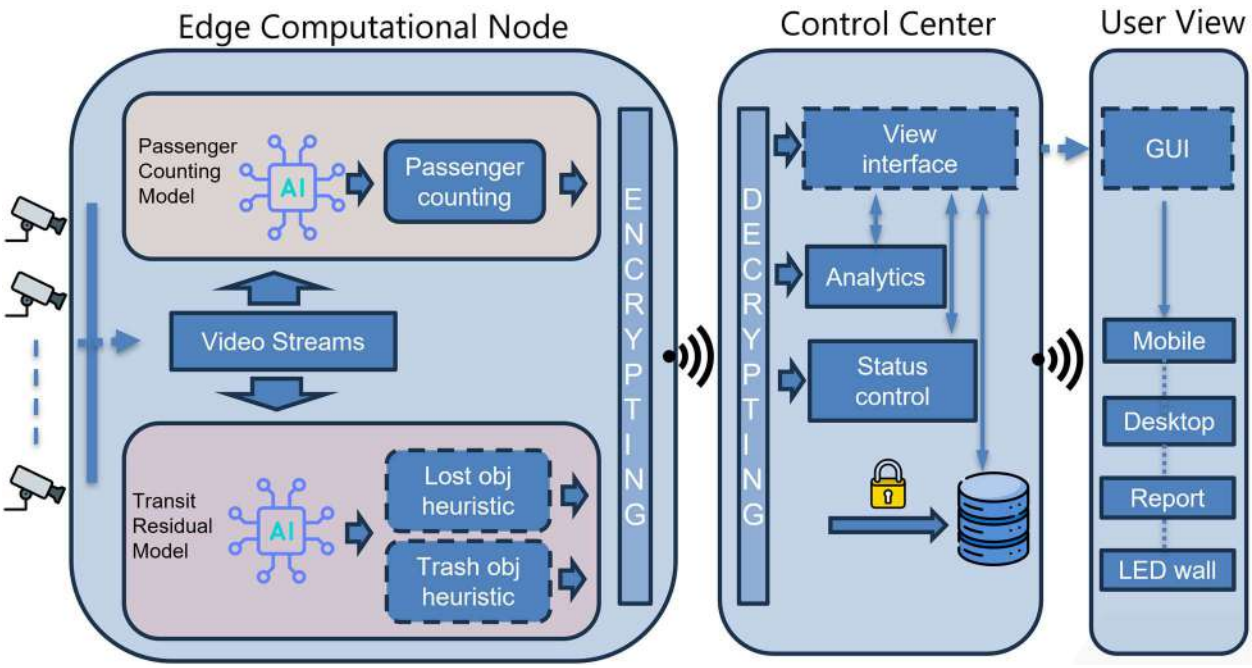

Figure 1 - The system architecture in the edge-oriented configuration.

During the initial phase of the project, the SUT was conceived to operate within a railway environment, leveraging onboard video streams acquired from train carriages during regular service. The architecture designed for the train environment was the one depicted in Figure 1.

However, as the project progressed, significant constraints emerged related to the authorization process required to access and process video data within the railway domain and install the system onboard. In particular, the involvement of the national Italian railway company Trenitalia and the stringent regulatory and operational requirements associated with railway systems led to delays that were incompatible with the project timeline. These limitations affected the ability to deploy, test, and iteratively refine the SUT in an operational railway setting.

To mitigate these risks and ensure the continuity and effectiveness of the experimental activities, a strategic decision was made to focus the alternative and comparable application domain within public transportation, which was already mentioned in the Description of the Action. The focus was therefore shifted from trains to buses, which offer a more accessible and flexible

environment for data acquisition and system deployment. This change significantly reduced the complexity of the authorization process and allowed for faster setup and validation cycles.

From a technical perspective, the most relevant change is the switch from an edge-oriented system architecture to a centralized one, where the AI SUT runs on a secure server machine.

However, the transition preserves the core characteristics and objectives of the original system. The new SUT continues to rely on video-based analytic and privacy-by-design principles. Moreover, the selected functionalities, such as passenger counting and detection of transit residual items, remain totally consistent with the initial choice, ensuring continuity in the evaluation framework and comparability of results.

In line with a privacy-preserving approach, raw video data are not subject to long-term storage for analytical purposes. The processing takes place in a controlled environment, and only anonymized and aggregated outputs are generated and retained. These outputs include non-personal information such as passenger counts, movement patterns within the vehicle, and detection of specific events related to objects.

The generated data is stored in a secure environment, with access regulated through authorization mechanisms to ensure compliance with applicable data protection requirements.

Within the pilot phase, the SUT operates under controlled and monitored conditions, including the selection of vehicles, configuration of sensors, and definition of operational scenarios. This setup allows for consistent data collection and supports iterative testing and progressive refinement of the system's functionalities.

Overall, the current configuration of the SUT enables the continuation of the experimental activities in a more accessible and manageable context, while preserving the initial objectives related to the evaluation of AI-based analytics in public transport systems.

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

2.1 Status of the 1st Pilot execution

As mentioned in section 1.3, the Initial pilot experimentation suffers from delays in the past year to obtain formal permissions by Trenitalia. Despite the agreement being signed, Trenitalia's internal bureaucracy and approval procedures lead to long waiting times even for the simplest tests that do not involve passengers. From September 2025, the pilot is running on alternative transport infrastructure (buses) which involve local and regional bus companies.

The activities done on the alternative platform are running fast, and we can say that the current TRLs are not far from what planned in the beginning, at least for the People Counting functionality (see Table 1).

Functionalities	Current/Expected TRL
Passengers Counting	6/7
Transit Residual Detection	3/5

Table 1 - Expected TRL of system functionalities.

The Transit Residual Detection accuracy results are not satisfactory yet (especially for garbage detection), we aim to mitigate this risk by developing a new transport related dataset, as it will be further explained in the following sections (see Section 2.1).

Two Stakeholder consultations have been done in 2025 as indicated in Table 2. Their feedback is reported and discussed in paragraph 2.1.2.

Workshop date	Number of participants
22 February 2025	38
6 June 2025	29

Table 2 - Stakeholders consultations in 2025.

The TrustSense and the TrustGuard tools have been used by the members of the team (AI developers from CNR and MERMEC Eng) after the training sessions held in February 2026.

Data and AI Model Passports for the AI Assets were successfully implemented on the FAITH AI Model Hub.

2.1.1 Deployment Context & Domain Specifics

The new SUT is deployed within a bus-based public transport environment, where video data are acquired through onboard CCTV systems installed on selected vehicles operating along a

predefined pilot line. Each vehicle is equipped with cameras positioned to monitor passenger areas, enabling the collection of visual data necessary for the execution of AI-based analytics. The video streams acquired onboard are recorded by a local Video Management System (VMS), and then they are sent via encrypted transmission to a controlled server environment as depicted in Figure 2.

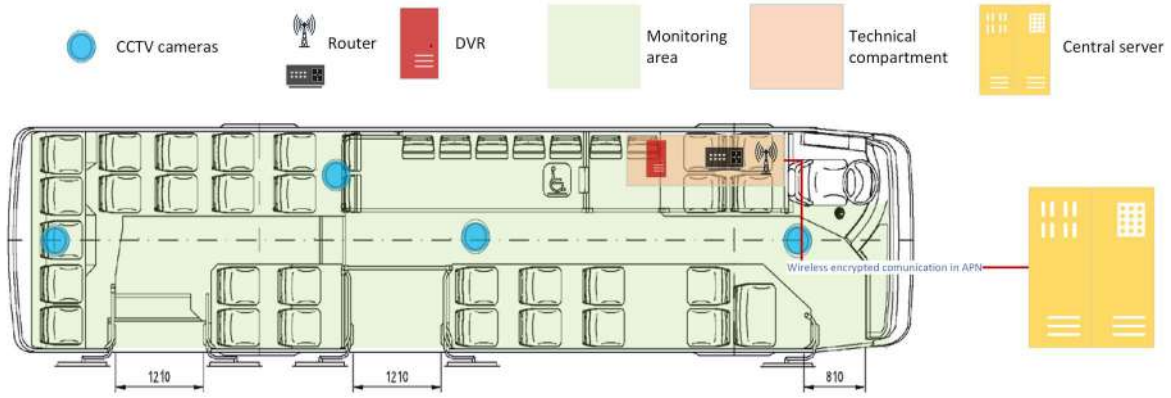

Figure 2 - The new experimental bus platform currently in use for the initial pilot

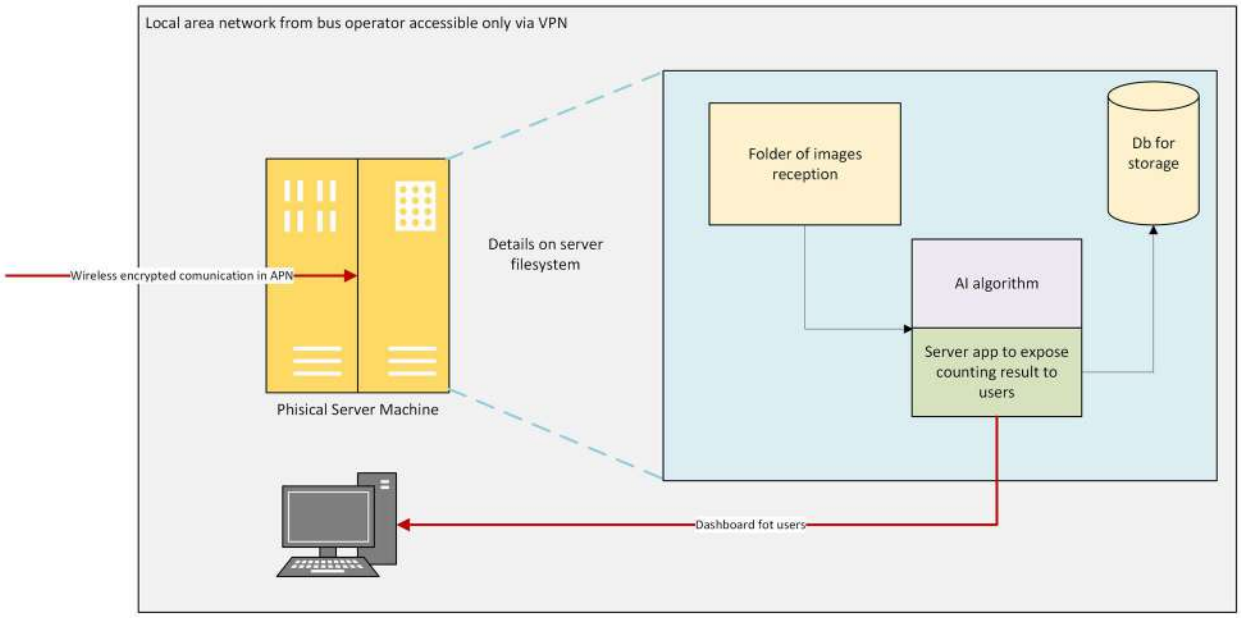

Figure 3 - The analytical functionalities are now executed at server level.

The analytical functionalities are executed at server level, within the operator's network infrastructure, allowing near real-time processing while maintaining centralized control over data

flows and system operations (see Figure 3). This architecture simplifies system management and ensures consistency in processing across the monitored fleet.

The pilot was carried out on a selected set of vehicles, including both urban and extra-urban buses, chosen to ensure sufficient similarity in terms of vehicle layout, camera positioning, and operating conditions. This allowed the testing activities to be conducted in a stable and comparable context, reducing variability not related to the performance of the system itself.

Within this setup, the system supported the extraction of non-personal transport indicators, such as passenger occupancy levels and other event-based outputs useful for service analysis. In line with the pilot nature of the activity, the deployment was carried out under controlled conditions, to support iterative validation and progressive refinement of the system in a real public transport scenario.

Passengers Counting

The Passenger Counting functionality is based on frame-by-frame object detection applied to video streams acquired from onboard CCTV cameras installed on buses operating in the pilot environment (see Figure 4). The purpose of this functionality is to estimate the number of passengers on board by identifying passenger silhouettes within the monitored areas of the vehicle.

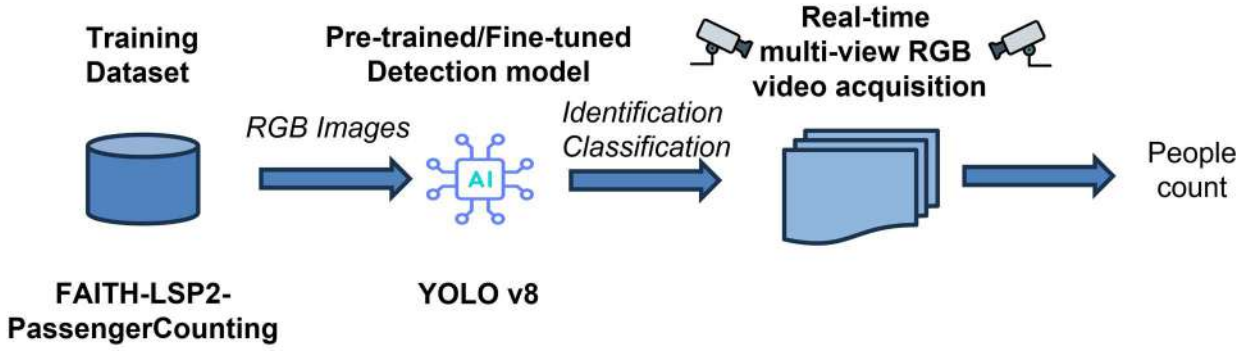

Figure 4 - The Passenger Counting functionality schema.

Although people's detection is a widely studied computer vision task, its direct application in the present pilot context cannot be considered fully reliable without domain-specific adaptation. Standard object detection approaches, when applied out-of-the-box, are affected by several operational factors that reduce performance in real public transport scenarios. In particular, the bus environment introduces specific challenges such as passenger occlusions, variable lighting conditions, non-homogeneous camera configurations, different layouts between urban and extra-urban vehicles, seated and standing postures, crowded conditions, reflections, vibrations, and oblique viewing angles.

For this reason, the Passenger Counting functionality requires adaptation and tuning to the target deployment domain. The pilot is intended to assess the extent to which the selected detection approach can operate effectively under these real operating conditions, and to identify the main sources of error affecting the counting results. In this respect, the pilot does not aim only at demonstrating technical feasibility, but also at supporting the progressive refinement of the system configuration for the specific transport context under analysis.

The processing pipeline is based on the extraction of individual frames from the video stream and on the application of an object detection model to identify passenger-related entities within each frame. Each detection is associated with a confidence score, and only detections above a predefined threshold are retained for counting purposes. In the current pilot configuration, the reference confidence threshold is set to 0.5, while the maximum processing time is approximately 50ms per frame.



Figure 5 - Passenger counting on board.

For passenger counting on board (Figure 5) after the doors closing, a fixed time is waited to allow passengers to distribute within the vehicle. Subsequently, a snapshot is captured and processed simultaneously from all cameras.



Figure 6 - Passengers boarding time.

For passengers boarding time (Figure 6) after the doors closing, for each door, the algorithm interprets passenger intent (internal (green)/external(red)) to avoid false counts caused by passengers standing at the doorway.

The validation of the Passenger Counting functionality is carried out through comparison with manually measured ground-truth data. For the pilot phase, the validation dataset was based on one selected vehicle, one operating day, and one reference trip used for detailed assessment. This setup was considered sufficient for an initial controlled evaluation of the SUT in the new domain, while also enabling the identification of the main environmental and configuration-related issues that may affect counting accuracy. Results on the validation activities are reported as a summary of values (Table 3) and a chart showing the trend in the difference with the ground truth (Figure 7).

Time	Manual	AI count	Error
27/03/2026 13:31:54	14	14	0
27/03/2026 13:37:05	19	19	0
27/03/2026 13:41:48	20	20	0
27/03/2026 13:49:43	44	40	+4
27/03/2026 13:51:29	60	53	+7
27/03/2026 13:52:42	59	55	+4
27/03/2026 13:55:28	55	52	+3
27/03/2026 14:05:08	31	32	-1
27/03/2026 14:06:11	30	30	0
27/03/2026 14:07:44	28	28	0
27/03/2026 14:09:04	25	25	0
27/03/2026 14:13:01	18	19	-1
27/03/2026 14:15:57	16	17	-1
27/03/2026 14:18:01	13	13	0
27/03/2026 14:21:58	7	7	0
27/03/2026 14:23:07	7	7	0
27/03/2026 14:26:02	12	12	0
27/03/2026 14:30:03	13	12	+1
27/03/2026 14:32:22	14	13	+1
27/03/2026 14:33:18	15	15	0
27/03/2026 14:36:24	17	17	0
27/03/2026 14:36:57	24	25	-1
27/03/2026 14:40:00	26	28	-2
27/03/2026 14:44:52	26	28	-2
27/03/2026 14:47:35	25	25	0
27/03/2026 14:50:40	32	34	-2
27/03/2026 14:52:32	36	31	+5

27/03/2026 14:54:45	32	28	+4
27/03/2026 14:56:38	38	33	+5
27/03/2026 14:57:40	39	35	+4
27/03/2026 15:03:26	30	29	+1
27/03/2026 15:06:03	23	22	+1

Table 3 - Data of the sample validation test.

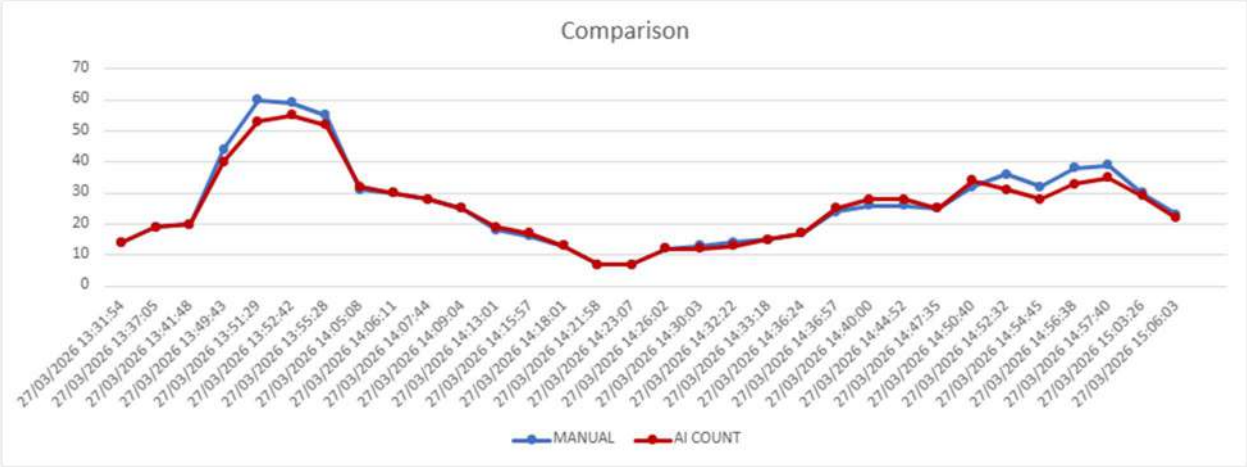

Figure 7 - The comparison of manual and AI count in the sample validation test.

The pilot results are therefore expected to provide evidence on the suitability of the selected approach for passenger counting in the bus domain, as well as indications for further tuning activities required to improve robustness and consistency under real operating conditions.

The system's performance was evaluated using the mean relative error, resulting in an average error of approximately 5.9% compared to manual counts. This corresponds to an overall accuracy of about 94.1%, indicating that the AI provides highly reliable quantitative estimates despite occasional deviations (see Table 4).

Vehicle type	Service type	Period	Accuracy	Measure Nr.
Urban vehicle	Interurban trip	27 March 2026	94.1%	32

Table 4 - Summary of the sample validation test.

Transit Residual Detection

The Transit Residual Detection functionality aims at identifying unattended items and garbage. This task is performed at the end of the service or even at the end of each run but always with the empty vehicle, i.e. no one is supposed to be present and recorded.

The identification of specific objects such as garbage and abandoned luggage (which are underrepresented in commonly used datasets such as *COCO*, *ImageNet*, etc.), cannot be effectively addressed using pre-trained models (e.g., *YOLO26n*[1]), which fail to achieve satisfactory performance against the objectives set by the Transportation LSP.

50 abandoned luggage items and 50 garbage elements were selected from those acquired. For each of these, we applied the state-of-the-art AI Model *YOLO26n* object detector. To handle the difference between the *YOLO26n* labels and those of the selected dataset, we modified the labels of the dataset according to those of *COCO*, thereby enhancing *YOLO26n's* ability to detect them.

Each of the 100 items was evaluated 72 times (once for each rotation), and the aggregated scores are reported in the table below.

Selected Dataset	Correct labelling	Sub-category labelling	Coherence between rotations	Q1-Q3 range
Garbage	18%	16%	53.4%	[25.3-74.3]
Luggage	40%	22%	49.8%	[27.1-70.8]

Table 5 - Out-of-the box Yolo26n performances in recognition of garbage and luggage items.

As expected, even state-of-the-art methodologies such as *YOLO26n* do not achieve acceptable performance in pre-fine tuning. They can determine the category of luggage approximately 40% of the time but fail to distinguish between luggage sub-categories 22% of the time (see Table 5). Additionally, even minor changes in view can result in significantly different outcomes, with an average consistency across views of only 49.8% of cases and a substantial interquartile range of [27.1–70.8]. This implies that the model is not suitable for identifying luggage without further fine-tuning, and that it is incapable of guaranteeing coherent labelling, even after slight modifications to the view of the item itself. The garbage identification task is even more complex because most existing databases do not label rubbish and prioritize models that avoid over-labelling rubbish and empty spaces. In conclusion, *YOLO26n* is inefficient at labelling rubbish, recognizing only cups and fruit, and maintaining suboptimal coherence between views (53.4% [25.3–74.3]).

[1] https://docs.ultralytics.com/models/yolo26/



Figure 8 - Performance issues of YOLO26n in the identification of garbage and luggage.

Further tests were performed on a small subset of ten internal camera views from the bus, including images of rubbish and chairs. While the model proves efficient at identifying chairs, it is less effective at identifying rubbish/garbage in the complex environment of a bus, resulting in poor performance that requires fine-tuning to achieve the desired level (see Figure 8 for some examples of failure). The sample size and manual labelling of the scenes are major bottlenecks in this analysis, which should be considered only as proof of the need to extend existing strategies.

BUS-Scenes	Correct labelling	Q1-Q3 range
Chairs	**48.9%**	**[0-81.2]**
Garbage	4%	[0-0.6]

Table 6 –Yolo26n performance are very poor on garbage detection in real scene

Therefore, to enhance the object detection capabilities of existing models by integrating label persistence strategies into the video stream, we decided to develop a dataset tailored to the task at hand.

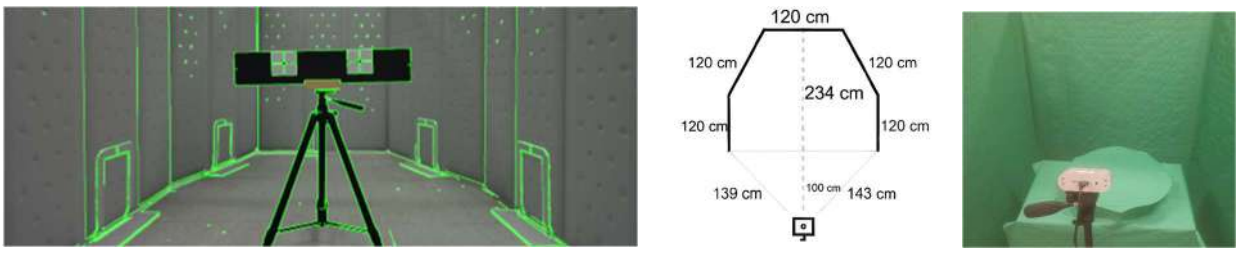

Figure 9 - Acquisition setup including rotating platform and adjustable artificial lighting.

By defining an appropriate controlled acquisition setup (Figure 9) with adjustable artificial lighting, a fixed-speed rotating platform for scientific acquisitions, and synchronization with a

multimodal camera (*ORBBEC FEMTO-BOLT*[2], RGB + depth mode, computed using infrared Time of Flight [*TOF*]), we proceeded with the multi-view and multi-modal acquisition of four types of objects: *luggage*, *dresses*, *confounding items*, and *garbage*. As represented in Figure 9, Panel a) shows the controlled setup with superimposed the detected boundaries used for the co-registration while panel b) shows the measures of the setup for replicability. Panel c) shows the relative positions of the camera and the rotating platform.

Each object is acquired in multiple views (i.e. from different camera positions and/or item dispositions), and 72 rotations of 5° provide a high level of variety to fine-tune and enhance existing methodologies (Figure 10).

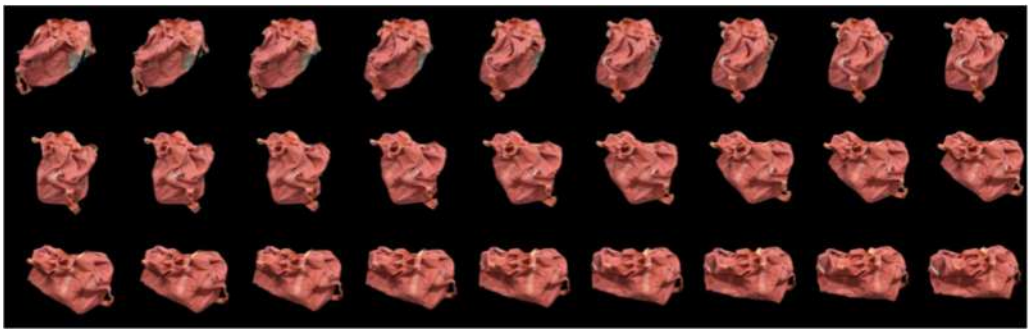

Figure 10 - Example of multiple views (72) for each item (e.g., red luggage in this image).

The acquired images were manually segmented using an open-source GUI developed specifically for this purpose (https://github.com/GDelCorso/SLImTAG), which incorporates state-of-the-art models (*Segment Anything Meta 3.1*[3]) to assist with segmentation using only a few ground-truth points. The software's co-registration of the RGB and depth fields, along with the automatic cleanup, results in over 300 items, each of which has between two and four views. This equates to over 50,000 labelled and segmented frames in total (Figure 11). Seventy-two multi-view acquisitions for each object were obtained using the ORBBEC FemtoBolt and co-registered across the depth and RGB streams. Semi-supervised cleanup was performed using SAM 3.1, followed by manual refinement.



Figure 11 - Multiview prototype ≈ 50000 images of 300 transit related items.

[2] https://www.orbbec.com/products/tof-camera/femto-bolt/
[3] https://ai.meta.com/blog/segment-anything-model-3/

Future development of the dataset includes the generation, using *generative AI* and *inpainting* methods (techniques already tested on appropriate subsets of the dataset, such as *LanPaint*[4] paired with *Juggernaut* models), of photorealistic scenes to be used for fine-tuning deep learning methods for the identification of trash and abandoned objects. Specifically, by capturing appropriate empty scenes (rooms, waiting rooms, platforms, interiors of public transportation vehicles, etc.) using the ORBBEC-FemtoBolt camera, the previously collected items will be placed within these scenes in a photorealistic manner, automatically inheriting their labels and enabling the creation of a dataset of 10,000–50,000 labelled scenes sufficient for the required fine-tuning.

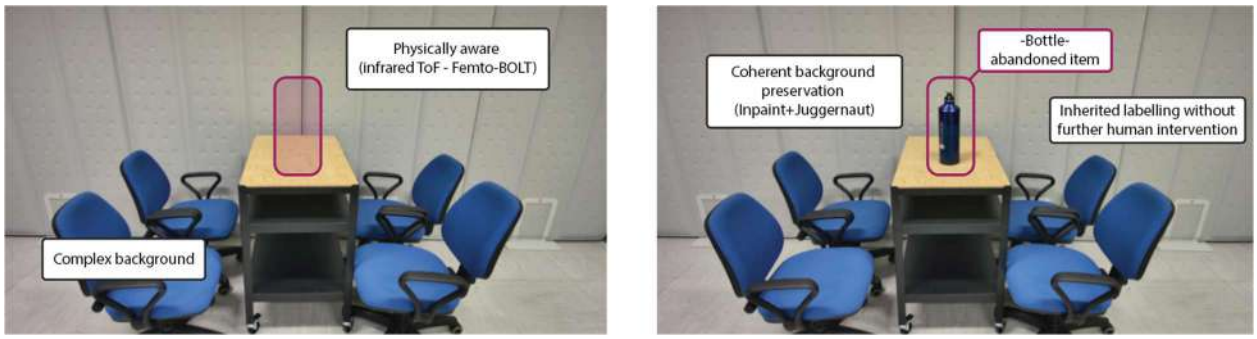

Figure 12 - Example of physically informed Generative image obtained by Inpaint pipeline.

2.1.2 Stakeholders workshop

To collect user feedback and to investigate the perceptions and expectations of potential users regarding the adoption of trustworthy AI systems in the railway sector, a targeted consultation was conducted with a sample of domain experts with more than 10 years of experience.

As the objective was not to maximize response rates but to gather informed and experience-based input, it was implemented through an online questionnaire administered during the DITECFER Acceleration Event on AI in the Rail and Mobility Value Chain[5], held in June 2025, Pisa (in Figure 13 the presentation of the FAITH project during the event).

[4] https://github.com/scraed/LanPaint

[5] https://www.ditecfer.eu/en/events/ditecfer-2nd-acceleration-event-ai-made-easy-for-the-railway-and-mobility-value-chain/



Figure 13 - The targeted expert consultation carried out.

Given this methodological approach, the results should be interpreted as indicative insights derived from a qualified expert sample, rather than statistically representative findings.

To ensure balanced representation in the feedback collected, respondents were drawn from both the private (51.7%) and public (48.3%) sectors. The sample included not only potential users in the railway domain, but also a relevant operator stakeholder active in an alternative bus platform.

Overall, 29 domain experts participated in the consultation, providing structured input based primarily on engineering (55.2%) and IT (51.7%) backgrounds, while also including managerial perspectives (17.2%). They represent organizations that, in most cases, already use (17.9%) or are planning to adopt (46.4%) assessment systems for AI-based technologies.

The consultation was specifically designed to capture user perceptions on barriers, opportunities, and strategic priorities for the integration of trustworthy AI in the transport domain.

With regard to perceived barriers, feedback consistently highlights the lack of transparency (51.7%) as a major obstacle to the adoption of trustworthy AI systems. More specifically, respondents emphasized the need for AI systems to clearly communicate their operational limitations and intended contexts of use.

User feedback also identified the most promising applications for the future of the sector, with three[6] areas standing out:

- Predictive maintenance optimization for both rolling stock and infrastructure (65.5%)
- Real-time detection of technical or safety anomalies (65.5%)

[6] Multiple-choice question, with a maximum of 3 answer options.

- Automated traffic control and dynamic circulation management (44.8%).
- Additional feedback indicated interest in passenger safety alert systems (24.1%) and AI-based tools for the automatic identification of waste in carriages, supporting on-board personnel (10.3%).

To support the broader adoption of trustworthy AI tools, platforms, and evaluation methods across the sector, potential users emphasized the need to address several key challenges, including:

- effective integration with legacy systems currently in use (48.3%);
- greater explainability of AI decision-making (41.4%);
- simplified processes for validation and certification of AI-based solutions (41.4%).

At the same time, feedback pointed to a strong and tangible willingness to experiment with such solutions as depicted in Figure 14. It is remarkable that less than one fifth (17.9%) of the organizations already use some assessment tools, but nearly two-thirds (60.7%) of the organizations represented expressed their willingness to test tools for assessing the trustworthiness of AI-based systems.

This indicates that the need for trustworthy AI systems is strongly felt and points to a significant market potential for AI solutions designed to ensure trustworthiness.

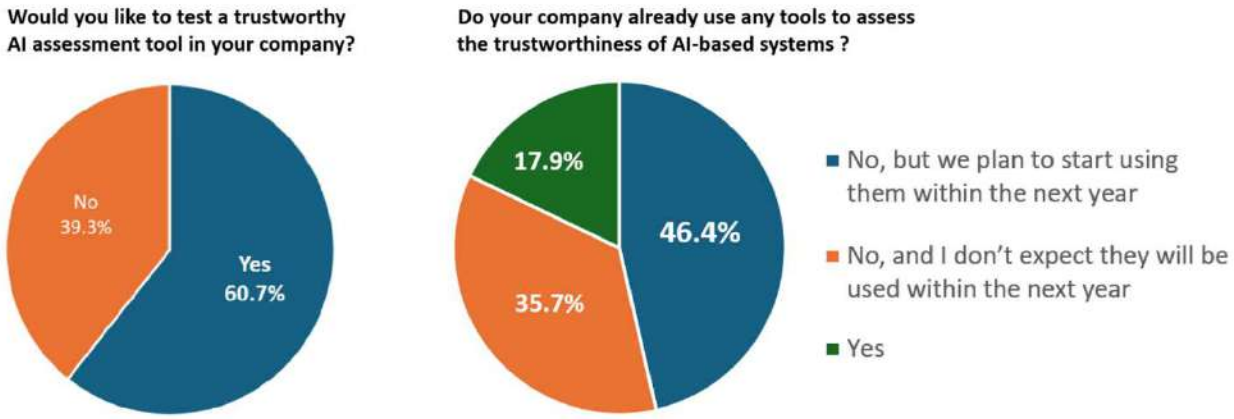

Figure 14 – Stakeholders' willingness to experiment with trustworthy AI solutions.

2.1.3 Users

Within the pilot, the main user groups are represented by domain experts and the AI technical team. These users interact with the SUT at different stages of deployment, validation, and performance assessment.

Domain experts include personnel with knowledge of the public transport service and of the operational context in which the pilot is carried out.

The AI technical team includes:

- technical AI operators and system administrators, responsible for system configuration, monitoring, management of the processing environment, and verification of the correct operation of the overall infrastructure;
- annotation and validation users, responsible for manual checks, data annotation, ground-truth preparation, and comparison between expected and generated outputs during the testing activities.

2.1.4 Training

User training is provided through dedicated sessions held either in person or via web conference. The training introduces the main functionalities of the SUT, including the available dashboards and the expected user interaction with the system.

Specific attention is given to the interpretation of the outputs, as well as to the main limitations of the system and the possible errors that may occur during operation. Example cases are presented to help users understand the experimental nature of the pilot and to promote informed and aware use of the tool.

For technical users, the training covered anomaly detection, maintenance activities, and verification of the correct functioning of the entire processing pipeline. It also included operational instructions for the correct system configuration in cases where vehicles are added, replaced, or modified during the pilot.

2.1.5 Evaluation Process

The evaluation process was carried out jointly with the customer during dedicated assessment sessions agreed according to the customer's operational availability. Evaluation activities were performed in correspondence with time periods and service conditions in which the transportation company normally conducts onboard passenger counting, either directly on the vehicle or through the review of recorded images.

During these sessions, the results produced by the SUT are compared with the corresponding reference observations collected by the company (Figure 15).

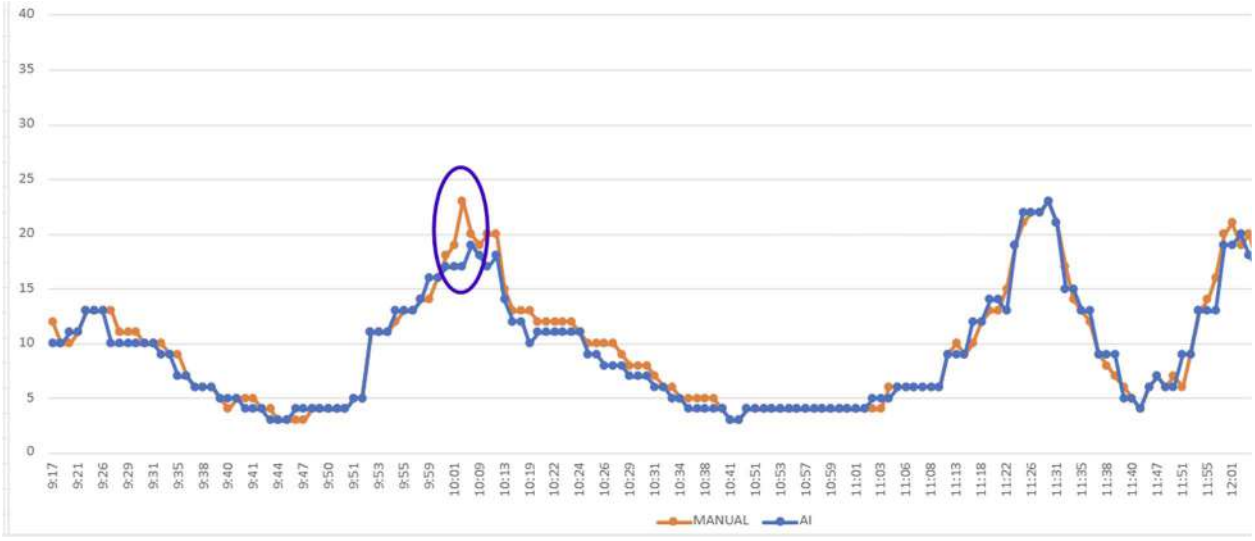

Figure 15 - The SUT results are compared with the ground truth observations.

This joint assessment allowed both parties to review the outputs together, discuss possible deviations, and develop a shared understanding of the pilot results and of the main factors affecting system performance. From the experimentation done during the pilot the most common sources of errors were camera occlusion (Figure 16) and sensor malfunctions (Figure 17).

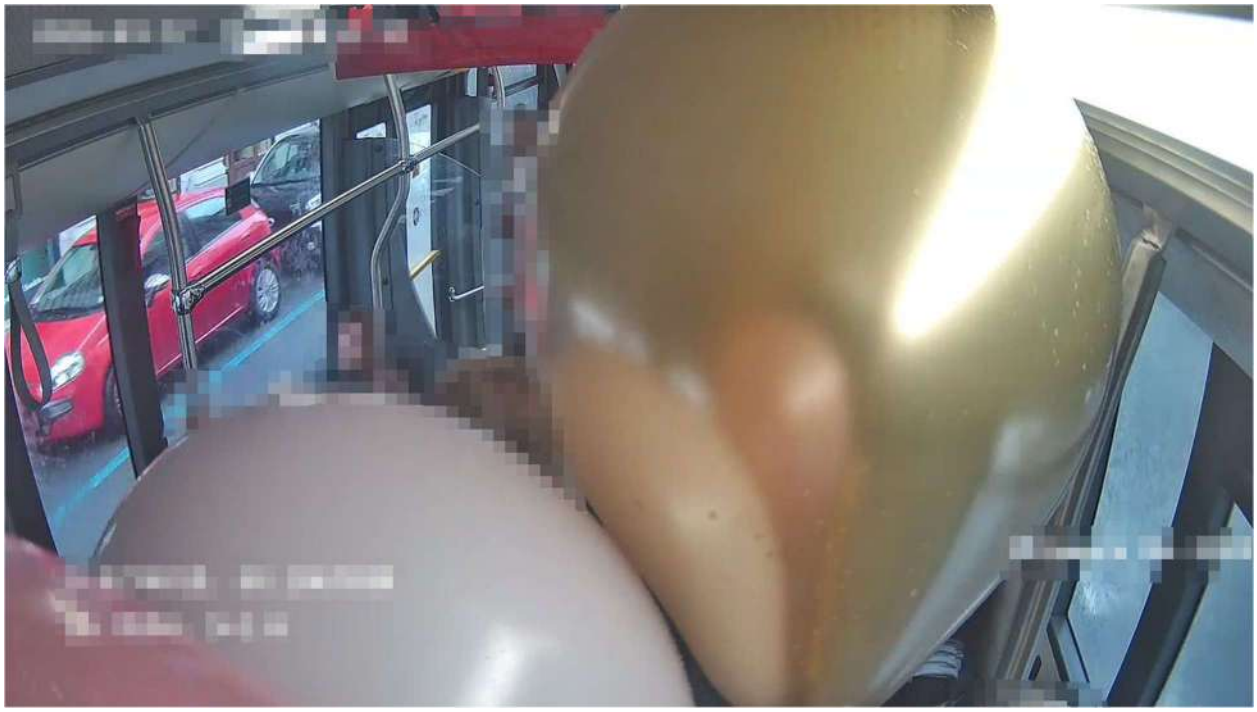

Figure 16 - Example of total occlusion of the camera view.



Figure 17 - Example of partial malfunctions of the camera sensor.

The evaluation process therefore combined technical verification with operational feedback, supporting a common interpretation of the test outcomes, and helping to identify possible areas for improvement.

2.2 Status of the Replication Pilot preparation/initialization

Contrary to the original plan, which required the system to be installed and tested on each specific train used for the trial, the alternative bus platform is already being installed across several Italian provinces. This ensures significant time savings during the Replication Pilot preparation phase, as it eliminates the need for initial hardware interventions and provides greater scalability than originally anticipated.

The future available fleet will consist of approximately *1,000 vehicles*, which encompasses all common categories of buses, specifically *Urban, Suburban, Intercity,* and *Bus Rapid Transit* models.

These vehicles are operated by several public transport companies across *10 Italian provinces:* Tper (Bologna, Ferrara) - SETA (Modena, Reggio Emilia, Piacenza) - Tep (Parma) - Atc esercizio (La Spezia) - START (Rimini, Ravenna, Cesena) - Amtab (Bari); Figure 18 depict the geographical position of the involved provinces; the blue ones are confirmed, while the violets, Bergamo and Pavia, should give an answer soon). The interactions had with the various local companies revealed that it is easier to negotiate with many local companies with respect to the big, centralised company of Trenitalia. In addition, these multiple contacts provide us with a larger number of stakeholders that can be involved in our consultation activities.

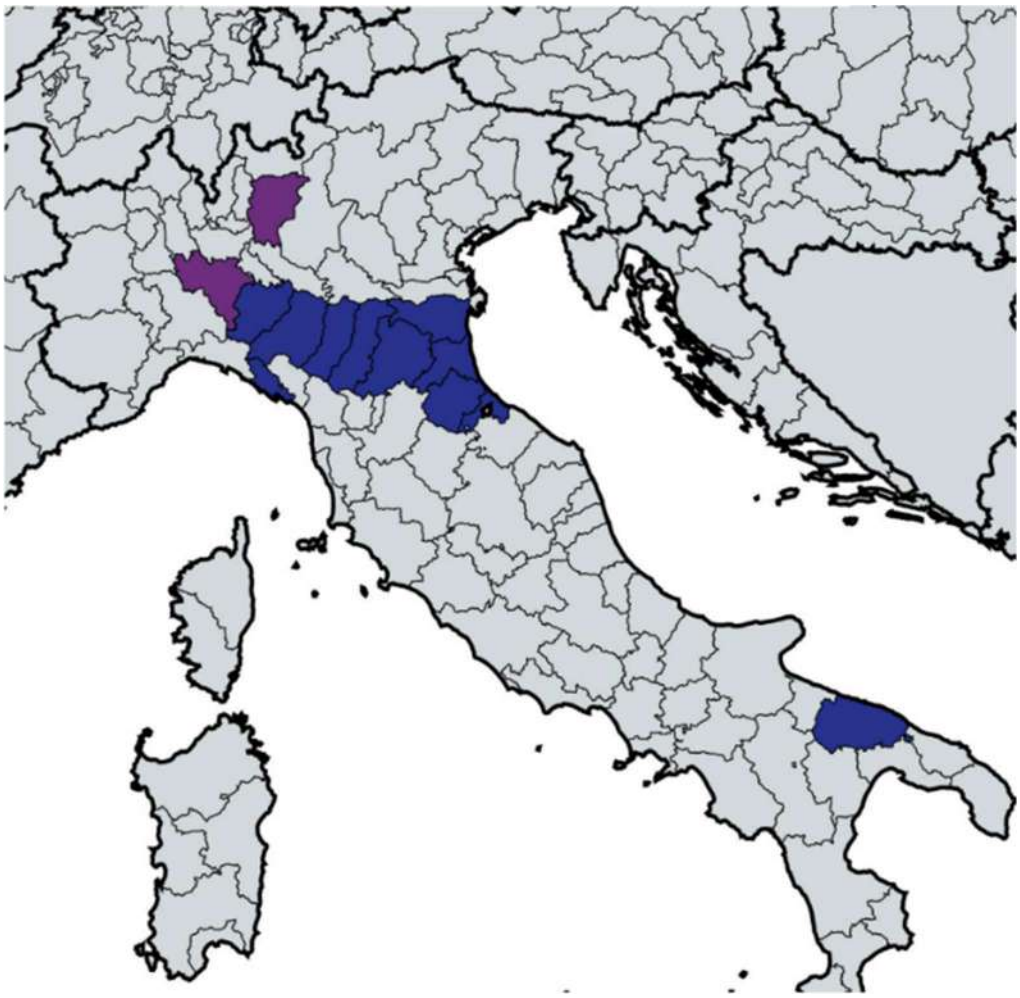

Figure 18 - The future fleet will consist of about 1000 vehicles in 12 Italian provinces.

In the Replication Pilot we will have extensive flexibility regarding the number and types of vehicles to be monitored, ensuring significantly greater scalability than originally planned.

2.3 Deviations from the original deployment plan

The LSP suffered from the delay in the first year to obtain formal permissions by Italian railways operator Trenitalia, to use their trains. Therefore, we presented an activity scheduled variation in Deliverable D5.1.

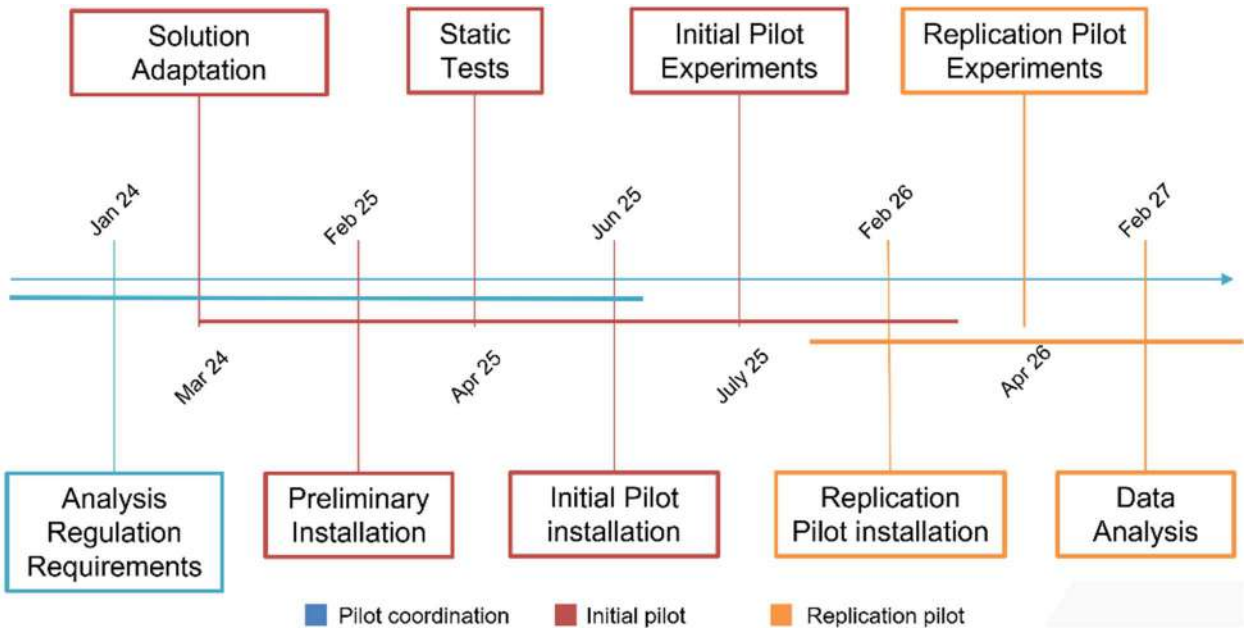

Figure 19 - Time schedule variation presented in Deliverable D5.1.

Despite the agreement being signed, Trenitalia's internal bureaucracy and approval procedures lead to long waiting times even for the simplest tests that do not involve passengers. From September 2025 the pilot is running on alternative transport infrastructure (buses), which involves local and regional bus companies. The initial pilot experimentation started in October 2025, 3 months later with respect to figure above.

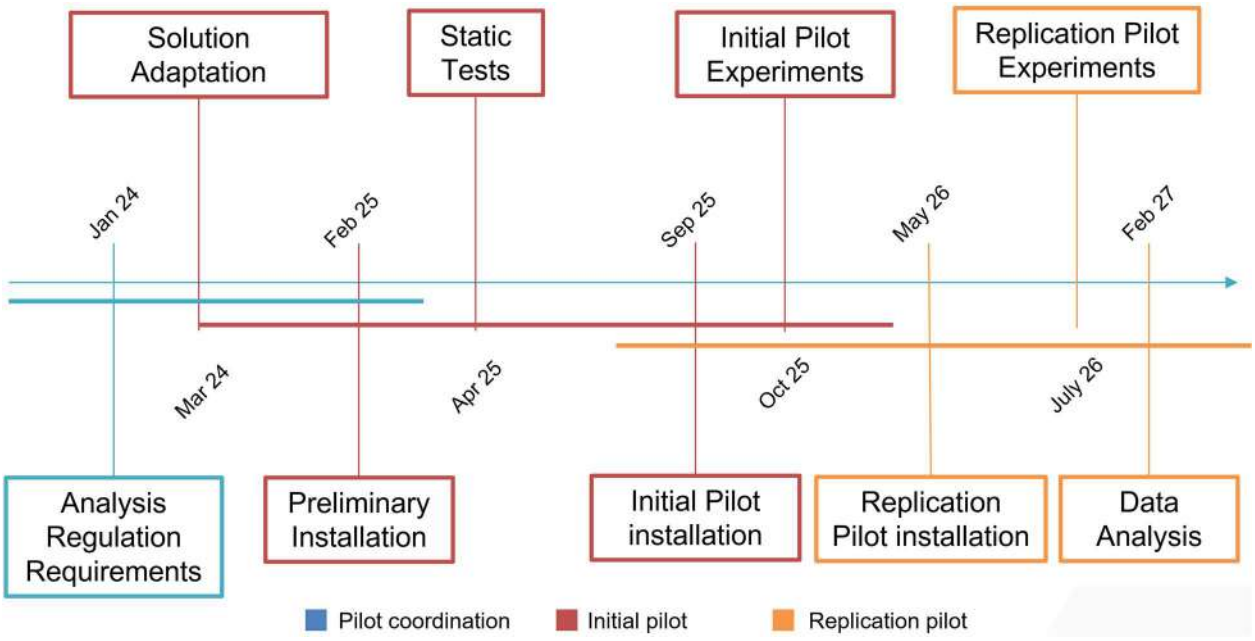

Figure 20 - New Time schedule variation.

With the replication pilot, we are confident to save a lot of installation time: the platform is already being activated in several Italian provinces, diverging from the original requirement for individual train-by-train installation. This avoids initial hardware overhead for the Replication Pilot, ensuring faster deployment and superior scalability compared to the initial project plan. We expect to mitigate the current three-month schedule slippage by the beginning of 2027 as shown in Figure 20.

2.4 Data collection throughout the 1st Pilot execution

During the 1st Pilot execution, data collection was carried out as a dedicated activity supporting the test and validation of the SUT in a real operational context. In the pilot, the purpose of data collection was not large-scale deployment, but the creation of a controlled set of reference cases that can be used to assess system behaviour together with the transportation company.

The data were collected from the onboard CCTV systems of the vehicles selected for the pilot and transferred through the existing infrastructure to the local server environment where the processing takes place. Throughout the pilot, the collected data were maintained within the customer's network and were handled exclusively inside the customer's infrastructure. The collection was planned on specific pilot runs and time intervals agreed for the testing activities, with particular attention to the situations that were most relevant for passenger counting, such as boarding, alighting, and changes in onboard occupancy along the route.

Within the pilot, the collected material was selected to include representative operating conditions, such as different passenger densities, seated and standing passengers, partial occlusions, and changes in lighting and camera viewpoint. This was necessary to ensure that the pilot evaluation reflects the actual conditions in which the SUT is expected to operate.

The collected data were then organized together with the main contextual information relevant to the pilot, such as vehicle, trip, timestamp, stop, and service conditions. A preliminary quality check was performed before the material was used for validation, in order to exclude incomplete or unsuitable data.

A subset of the collected material was then annotated in CVAT[7] to create the pilot reference dataset (see Figure 21). Through this activity, authorized personnel reviewed the images and video sequences and produced the ground-truth needed for validation. Manual counting and verification may also be performed either onboard the vehicle or through the review of recorded images, depending on the selected pilot session. The annotated dataset was then used for comparison with the AI Model outputs and for the joint analysis of the pilot results together with the transportation company.

Overall, data collection in the 1st Pilot execution was structured to support a practical and shared assessment of the SUT, based on real service conditions, controlled acquisition, and verifiable

[7] https://www.cvat.ai/

reference data, while keeping the data within the customer's network environment throughout the pilot.

Down below an example of folder structure for AI training:

```
dataset/
|
├── images/
|  ├── train/
| |  ├── img001.jpg
| |  ├── img002.jpg
| | └── ...
| |
|  ├── val/
| |  ├── img101.jpg
| |  ├── img102.jpg
| | └── ...
| |
|  └── test/
|  ├── img201.jpg
|  ├── img202.jpg
|  └── ...
|
├── labels/
|  ├── train/
| |  ├── img001.txt
| |  ├── img002.txt
| | └── ...
| |
|  ├── val/
| |  ├── img101.txt
| |  ├── img102.txt
```

```
| | └── ...
| |
| └── test/
| ├── img201.txt
| ├── img202.txt
| └── ...
|
└── data.yaml
```

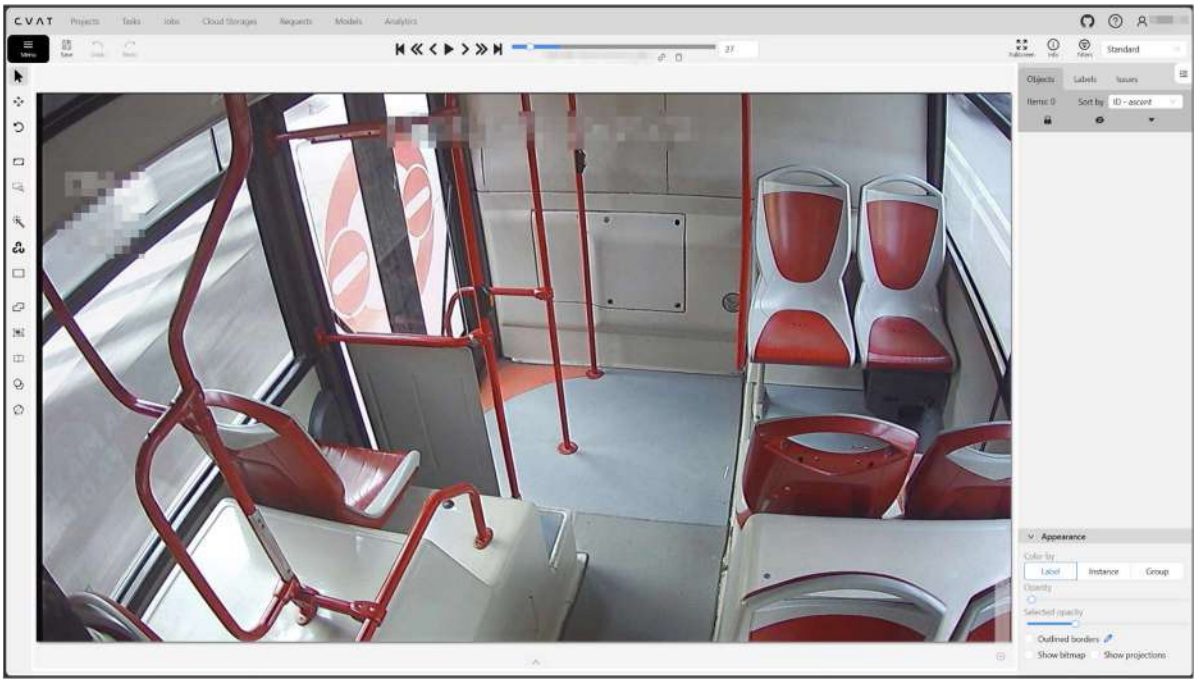

Figure 21 - Images of the reference dataset are annotated using CVAT

2.5 Integration into LSP workflows: Lessons learnt

The initial pilot provided valuable insights into how the SUT can be integrated into the larger-scale workflows of the transport service provider. While the pilot aimed at testing specific functionalities in a controlled setting, it also highlighted several challenges and areas for improvement when moving towards full-scale deployment.

2.5.1 Issues encountered and solutions

1. Data Quality and Collection Challenges. One of the primary challenges encountered during the pilot was ensuring the consistency and quality of the data collected from the onboard CCTV systems. Variations in lighting, passenger density, and camera angles led to occasional inaccuracies in the passenger counting process. To address this, additional training was provided to the technical staff responsible for camera setup, along with enhanced monitoring during data collection to identify and resolve potential issues in real-time.

2. System Configuration and Calibration. Another challenge was related to the configuration and calibration of the SUT when changes in operational conditions occurred. This highlighted the need for flexible, well-documented procedures to ensure the system can be easily adapted to new vehicles.

3. Interactions with Users. The interactions with end-users (i.e., technical operators, system administrators, and transport planners) were critical in identifying potential usability issues and understanding the system's real-world performance. While the system generally performed well, the users experienced some difficulties in interpreting the outputs. To improve this, training sessions were scheduled, with a focus on how to interpret the system's results under various conditions.

2.5.2 Lessons Learned for LSP Workflow Integration

The integration of the pilot results into the larger LSP workflow provided several key lessons for full-scale deployments:

- Collaborative Approach with End-Users: continuous collaboration with users throughout the pilot phase was crucial to refine the system and ensure that its outputs aligned with operational needs.
- Scalable and Flexible System Configuration: ensuring that the system can adapt to a variety of vehicles, service types, and operational scenarios is key for scalability. The pilot emphasized the importance of establishing clear, standardized procedures for system configuration and calibration that could be easily replicated when the system is deployed on a larger scale.
- Real-Time Monitoring and Data Quality Management: the pilot highlighted the importance of monitoring to ensure data quality.

3 Application of FAITH Tools

3.1 Application of the FAITH AI TAF V1 methodology

The application of the FAITH AI Trustworthiness Assessment Framework (AI_TAF v1) within the Transportation LSP followed a structured, phase-based approach for assessing the trustworthiness of AI systems and their associated human actors. The phases have been extensively presented into Deliverable D2.1 and updated in D2.2. Each phase incrementally refines understanding of risks and trustworthiness characteristics for the asset under evaluation. In the context of the Transportation LSP as in the other LSPs, this workflow has been operationalized through the FAITH toolset, including TrustGuard, TrustSense, and the FAITH Model and Data Passports, which support the different aspects of the assessment lifecycle.

The following sections summarize the activities done and the results obtained.

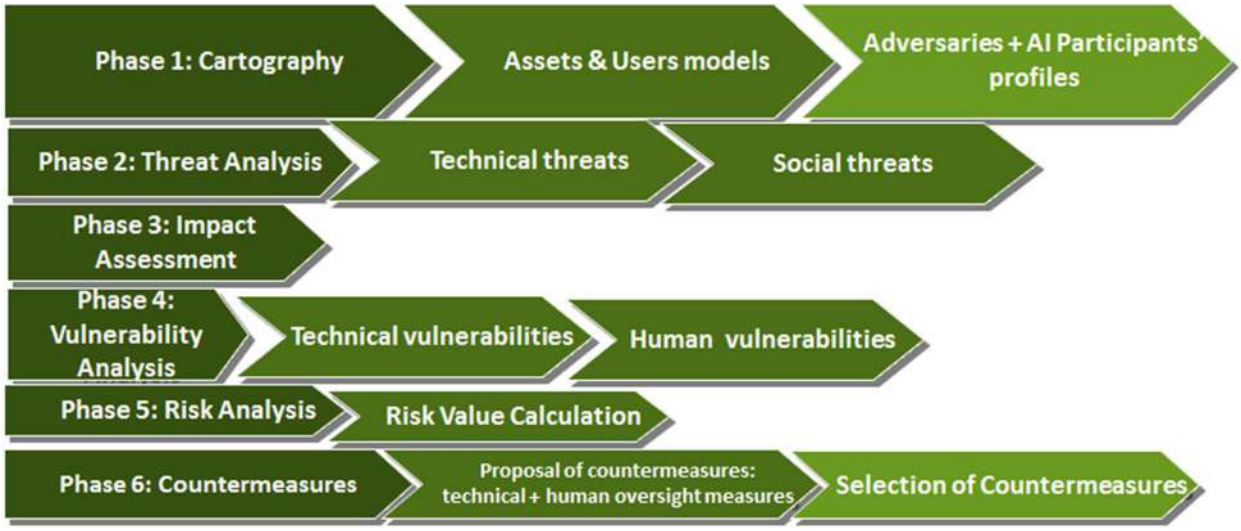

Figure 22 - The FAITH TAF V1 methodology.

3.1.1 Cartography

The Cartography phase consisted of a clear identification of the core SUT assets and their characterization. For the Transportation LSP, we have two main AI models, one for passenger counting, M1, and one for residual (or abandoned) object detection, M2. These have been trained and evaluated on two different datasets, which have been split into training (Dtr1 and Dtr2) and test (Dte1 and Dte2).

The following Table 7 summarizes the assets and their main characteristics.

AI Asset type	AI Asset name	Lifecycle stage	Team owner per AI Asset	AI team maturity	Comment
AI ASSET - Models	M1-LSP2-Passenger Counting	DEPLOY	MERMEC	High	
AI ASSET - Models	M2-LSP2-Transit Residual Detection	DEVELOP	CNR	High	
AI ASSET - Datasets	Dtr1-FAITH-LSP2-Passenger Counting	DEPLOY	MERMEC	High	70% of the annotated dataset is used for training
AI ASSET - Datasets	Dte1-FAITH-LSP2-Passenger Counting Eval	DEPLOY	MERMEC	High	30% of the annotated dataset is used for evaluation
AI ASSET - Datasets	Dtr2-LSP2-Transit Residual Dataset	DEVELOP	CNR	High	70% of the annotated dataset is used for training
AI ASSET - Datasets	Dte2-LSP2-Transit Residual Dataset Eval	DEVELOP	CNR	High	30% of the annotated dataset is used for evaluation

Table 7 - List of the SUT assets.

The criticality levels based on an initial classification carried can be summarized as reported in Table 8 below, showing that the only relevant criticalities are (i) the AI system poses low or no risk of harm to the health, safety or fundamental rights of natural persons, including influencing the outcome of decision making; and (ii) since the AI system processes personal data, GDPR does apply, but since video streams are not even recorded, the criticality level is assessed as medium to low risk.

Criticality assessment			
Questions	Useful resources	Answer	Comment
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?	AI Act Annex III[8]	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? 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[9]	No	
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 Directive	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?	Annex 2 of NIS2 Directive	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?		Yes	
Is the AI system processing personal data? Does GDPR apply?	GDPR[10]	Yes	Medium to low risk (Video streams are not even recorded)

Table 8 - AI Act classification of criticality levels.

[8] https://artificialintelligenceact.eu/annex/3
[9] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022L2555
[10] https://eur-lex.europa.eu/eli/reg/2016/679/oj?eliuri=eli%3Areg%3A2016%3A679%3Aoj&locale=en

This evaluation is in line with the results obtained from the application of the TrustGuard, as reported in Subsection 3.3.3.

3.1.2 Threat Analysis

Initial analyses related to possible threads were carried out considering the need to prepare a Data Protection Impact Analysis (DPIA) with the transportation company. This activity was organized to be overall in line also with the FAITH AI TAF V1 methodology and to provide feedback to the dry-run activities. To this end, the Threat Analysis phase involved systematically identifying and characterizing potential threats to the assets under consideration: AI models, datasets, system infrastructure, and human actors. At this stage, the analysis was conducted at the asset level without considering the full socio-technical system in context, in accordance with the phased approach defined in FAITH. This means that each individual component of the system, both the datasets and both the AI models assets (e.g., Passenger Counting and Transit Residual Detection) and the human assets (e.g., developers and operators interacting with the system, AI developers, data analysts) were assessed in isolation.

The SUTs process video streams acquired from on-board cameras in public transport environments. This operational context introduces a broad spectrum of threats spanning technical, organizational, and ethical dimensions.

The threat identification process has been initially guided by the DPIA-related considerations and it was done in line with recognized risk management frameworks, including ENISA Risk Assessment, ISO/IEC 23894:2023 (Artificial Intelligence — Risk Management), the NIST AI Risk Management Framework (AI RMF 1.0) and the EU Artificial Intelligence Act's requirements for AI systems, particularly with regard to robustness, data governance, transparency and human oversight. Afterwards the identified threats have been mapped to FAITH Golden Table and added if not present. Desk analyses have been also conducted to better understand instantiations to the specific transportation domain.

From a technical perspective, the following threats have been considered:

- **Adversarial manipulation of AI models** involves both intentional and unintentional alterations to input data that can affect the behaviour of computer-vision algorithms. In the context of the SUT, such threats may arise from adversarial examples (e.g. deliberately crafted visual patterns), environmental perturbations (e.g. changes in lighting, occlusions and motion blur) and physical manipulation of objects in the real world. Such threats could lead to inaccurate passenger counting, missed detections of unattended objects, and inconsistent results across frames. This category of risk is directly related to the robustness requirements defined in the EU AI Act and is addressed in both ISO/IEC 23894 and the NIST AI RMF.
- **Data poisoning and data integrity risks** arise from the potential contamination of training, validation, or fine-tuning datasets. These risks are particularly relevant due to the development of custom datasets and the use of synthetic data generation techniques. Sources of threat include annotation errors, unbalanced datasets, artefacts in synthetic

data, and inconsistencies in the data pipeline. These risks may result in reduced model accuracy, poor generalization and biased outputs. They are directly linked to the data governance requirements set out in Article 10 of the EU AI Act and are also covered by ISO/IEC 23894 and the NIST AI RMF.

- **Model performance degradation (model drift)**, due to discrepancies between controlled training environments and real-world operational conditions, such as variations in lighting, crowd density, and camera positioning. This can lead to reduced reliability over time and necessitates continuous monitoring and validation, as outlined in ISO/IEC 23894.
- **Sensor and hardware vulnerabilities**, including camera malfunctions, occlusions, or misalignment between RGB and depth data streams, potentially affecting detection accuracy and system reliability.
- **Integration and system-level vulnerabilities**, such as communication failures between edge devices and backend systems, latency issues, or data synchronization errors.
- From a data protection and ethical perspective, the following threats have been identified:
- **Privacy risks**, including potential re-identification or unintended processing of personal data. These risks are closely linked to EU AI Act requirements and GDPR principles.
- **Lack of transparency and explainability**, identified by stakeholders as a major barrier to adoption, potentially limiting trust and operational acceptance. This aligns with transparency obligations under the EU AI Act and NIST AI RMF principles.
- **Bias and unfairness in AI outputs**, potentially arising from dataset limitations or imbalanced representation of real-world scenarios, affecting the reliability and fairness of system outcomes.
- **Human-AI interaction risks**, related to how operators interpret and act upon AI outputs. These include over-reliance (automation bias), mistrust, misinterpretation of outputs, and usability issues. Such risks may lead to incorrect or delayed responses in safety-critical situations and are explicitly addressed in the EU AI Act (Article 14), as well as in ISO/IEC 23894 and the NIST AI RMF.

From an organizational and operational perspective, the following threats have been considered:

- **Dependency on external stakeholders and data providers**, impacting system deployment and data availability;
- **Insufficient user training and awareness**, potentially leading to misuse or incorrect interpretation of system outputs.

For a concise visualization of the identified threats, including the estimation of their frequencies, see Table 9, which also reports their impact and severity. These analyses, done for preparing the DPIA with the transportation company, have guided the dry-run assessment entailing a mapping of these threads to those included in the FAITH Golden Table, with initial considerations also for impact assessment.

Threat / Risk Source	Primary Impact	Impact Dimension	Severity qualitative)	Frequency
Adversarial manipulation of inputs	Incorrect detections and missed events	Technical / Safety	High	Very Rare
Environmental perturbations	Reduced accuracy and unstable outputs	Technical / Operational	Medium–High	Frequent
Data poisoning/poor data quality	Biased or unreliable model behaviour	Technical / Ethical	High	Rare
Model drift	Performance degradation over time	Technical / Operational	Medium–High	Rare
Sensor / hardware failures	Loss or corruption of input data	Technical / Operational	Medium	Occasional
Integration / communication failures	Data loss, latency, or inconsistent system behaviour	Technical / Operational	Medium	Occasional
Privacy risks	Legal and reputational consequences	Legal / Ethical	High	Rare
Lack of transparency / explainability	Reduced user trust and adoption	Human / Organizational	Medium–High	Frequent
Bias and unfairness	Unequal or unreliable outcomes	Ethical / Social	Medium–High	Occasional
Human-AI interaction risks	Incorrect or delayed operational decisions	Human / Safety	High	Occasional
Dependency on external stakeholders	Deployment delays and limited data availability	Organizational	Medium	Frequent
Insufficient user training	Misuse or misinterpretation of system outputs	Human / Operational	Medium–High	Rare

Table 9 - Mapping between identified threats, associated impacts and key dimensions.

Frequency levels are defined qualitatively as: Very Rare (exceptional), Rare (unlikely but possible), Occasional (sporadic), Frequent (regular), Very Frequent (systematic). These levels were then mapped onto the FAITH Golden Table frequency as follows: very rare → never; rare → once every 10 years; occasional → once every two years; frequent → once a year; very frequent → twice a year.

Following the domain-specific identification in Table 9, the Transportation LSP threats were aligned with the FAITH Golden Table to reuse a consistent taxonomy of threats, vulnerabilities and controls and to enable the dry-run assessment via questionnaire-based checks on controls. The mapping was performed at asset level by associating each identified threat to the closest Golden Table entry; any residual domain-specific nuances were captured as comments in the dry-run spreadsheet. The resulting mapped set is reported in Table 11 based on ENISA report [1]. The

mapping from the threat labels used in Table 9 (first column) to the threat labels used in Table 11 (third column) is reported in Table 10. Where a Table 9 entry corresponds to more than one Golden Table threat, all relevant Table 11 labels are listed.

Table 9 – Threat / Risk Source	Threat from FAITH Golden Table)
Adversarial manipulation of inputs	Evasion (adversarial examples crafted in black box conditions); Use of adversarial examples crafted in white or grey box conditions
Environmental perturbations	Failure or malfunction of ML application
Data poisoning/poor data quality	Poisoning; Label modification
Model drift	Failure or malfunction of ML application
Sensor / hardware failures	Failure or malfunction of ML application
Integration / communication failures	Failure or malfunction of ML application
Privacy risks	Data disclosure; Model disclosure; Model or data disclosure
Lack of transparency / explainability	Human error; Cybersecurity incident not reported to incident response teams
Bias and unfairness	Label modification
Human-AI interaction risks	Human error
Dependency on external stakeholders	Cybersecurity incident not reported to incident response teams
Insufficient user training	Human error

Table 10. Mapping between DPIA threats listed in Table 9 and the threats in FAITH Golden Table

Lifecycle Stage	Affected AI Assets	Threat	Question	Answer	Comment
Design	Data; Model	Label modification	How often do attackers in similar AI systems modify the training data in this sector?	Never	Models will not be retrained with data acquired on board. Actually, data acquired on board will be deleted soon after the model inference.
Testing	Model; IT Systems Network and Computing	Evasion (adversarial examples crafted in black box conditions)	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 and observe output.

					CCTV is a closed system, and such attacks are rare because the output that can be observed are only the alert
Testing	Model; IT Systems Network and Computing	Use of adversarial examples crafted in white or grey box conditions	How often do attacks with adversarial examples crafted in white or grey box conditions occur in this sector?	Never	In the designed system the attacker cannot have access to information (models, model parameters, etc)
Testing	Model; IT Systems Network and Computing; Maintenance	Oracle	How often do oracle attacks occur in similar AI systems in this sector?	Never	There are no observable outputs if you don't have access to the system.
Production	Data; Model; IT Systems Network and Computing; Maintenance	Failure or malfunction of ML application	How often do such ML applications fail or malfunction in this sector?	Twice per year	Image quality degradation (still frames, damaged pixels), Intentional occlusions and involuntary camera interference/occlusions (dirt, condensation, cobwebs). This will be estimated after proper observation.
Production	Data; Model; IT Systems Network and Computing	Denial of service due to inconsistent data or a sponge example	How often does denial of service occur due to inconsistent data or a sponge example in similar AI	Never	Inconsistent data belongs to the previous threat: failure or malfunction

			systems in this sector?		
Design; Testing; Production	Data; Model; IT Systems Network and Computing; Maintenance	Poisoning	How often do poisoning attacks occur in similar AI systems in this sector?	Never	Related to a specific vulnerability: data or model access protection
Design; Testing; Production	Data; Model; IT Systems Network and Computing; Maintenance	Model or data disclosure	How often do the models used in this sector disclose information about their configuration ?	Once every 10 years	
Design; Testing; Production	Data; Model; IT Systems Network and Computing	Data disclosure	How often are the data manipulated by the ML algorithms leaked in this sector?	Once every 10 years	
Design; Testing; Production	Model; IT Systems Network and Computing	Model disclosure	How often are the model internals (i.e., parameter values) of the AI system leaked in this sector?	Once every 10 years	Models are in well protected proprietary systems; very rare leakage possibility
Design; Testing; Production	Model; IT Systems Network and Computing; Maintenance	Compromise of ML application components	How often are the components of similar ML applications compromised in this sector?	Once every 10 years	Component compromise is rare but possible, especially if third-party components are used.

Design; Testing; Production	Model; IT Systems Network and Computing; Maintenance	Human error	How often do such ML applications fail or malfunction due to end-user error in this sector?	Twice a year	End-user errors can occur, especially in complex systems, leading to occasional failures or malfunctions.
Design; Testing; Production	IT Systems Network and Computing; Maintenance	Cybersecurity incident not reported to incident response teams	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 sometimes happens.

Table 11 - Threat frequencies

3.1.3 Impact Assessment

Again, an initial impact assessment was guided by the needs of preparing the DPIA with the transportation company but was also useful when contributing to the dry-run of FAITH tools.

The analysis considered multiple impact dimensions, including technical performance, operational effectiveness, safety, privacy, ethical aspects, and overall system trustworthiness.

From a technical and operational perspective, the main potential impacts include:

- **Degradation of system performance**, resulting from adversarial perturbations, data quality issues, or model drift. This may lead to reduced accuracy in passenger counting and object detection, directly affecting the reliability of the system.
- **Inconsistent or unstable AI outputs**, particularly under varying environmental conditions (e.g., lighting changes, occlusions, crowd density), which may undermine confidence in the system and limit its operational usability.
- **Failure to detect critical events**, such as unattended objects or hazardous conditions (e.g., garbage accumulation), potentially impacting safety and service quality in public transportation environments.
- **Error propagation across system components**, where inaccuracies in upstream detection tasks affect downstream processing, aggregation, or decision-support functionalities.

From a safety and human-related perspective, the following impacts have been identified:

- **Incorrect or delayed operational decisions**, resulting from misleading, incomplete, or ambiguous AI outputs in safety-relevant contexts;

- **Over-reliance or under-reliance on AI systems**, leading respectively to automation bias or disregard of relevant alerts;
- **Increased cognitive load for operators**, especially when system outputs are inconsistent or not sufficiently interpretable;
- **Reduced trust and acceptance of the system**, particularly in cases where AI behaviour is not transparent or explainable.

These aspects are directly related to human oversight requirements defined in Article 14 of the EU AI Act and are further emphasized in ISO/IEC 23894 and the NIST AI RMF, which highlight the importance of human-centred AI design and clear allocation of responsibilities.

From a data protection and ethical perspective, the following impacts have been considered:

- **Potential privacy violations**, in case of unintended processing or re-identification of individuals, leading to legal, ethical, and reputational consequences;
- **Bias and unfair outcomes**, resulting from dataset limitations or imbalanced data distributions, affecting the reliability and equity of system outputs;
- **Lack of transparency and accountability**, which may hinder regulatory compliance and reduce stakeholder trust.

These impacts are aligned with the requirements of the EU AI Act regarding data governance (Article 10), transparency (Articles 13–15), and protection of fundamental rights.

From an organizational perspective, the following impacts have been identified:

- **Delays in deployment and operational constraints**, due to dependency on external stakeholders or infrastructure limitations;
- **Increased operational and maintenance costs**, associated with continuous monitoring, retraining, and validation of AI models;
- **Challenges in certification and regulatory compliance**, particularly for AI systems, potentially classified as high-risk under the EU AI Act.

These considerations contributed to the content of the impact assessment in terms of the organizational or business consequences for the various threats and vulnerabilities, as summarized by the following Table 12. Overall, the identified impacts highlight the importance of adopting a comprehensive and risk-based approach to AI system design and deployment, ensuring that technical performance, human factors, and regulatory requirements are jointly addressed.

Trustworthiness characteristic	Question	Answer	Please provide in detail potential business and legal consequences
Safety	What is the business consequence level if the safe operation of	Negligible	The system is not designed to deal with safety threatening situations.

	the AI system is compromised?		
Security and resilience	What is the business consequence level if the security or the resilience of the AI system to adverse events is compromised?	High	Carriage maintenance is not always easy and frequent enough to avoid or recover from unfavourable events (e.g. spray on cameras). This could lead to the service interruption and loss of trust for users.
Explainability and interpretability	What is the business consequence level if the AI system is not interpretable?	Moderate	Interpretability aids in trusting the system's decisions, it is a negligible aspect when the accuracy values are high enough. On the other hand, providing, e.g., the end-user an image showing the region causing the alert, may be a valuable feature, which could increase a lot the trustworthiness in the user.
	What is the business consequence level if the AI system is not explainable?	Moderate	Engineers need to understand why alarms or triggers are flagged to plan countermeasures; lack of explainability slows root-cause analysis.
Privacy	What is the business consequence level caused by a privacy compromise of the AI system?	Moderate	(DPIA) The severity of the risk is considered medium because illegitimate access to data could cause some possible damage to the data subjects, compromising their privacy and potentially generating negative impacts on their reputation, especially if interpreted out of context. In fact, the data acquired in DEV mode may include both legitimate behaviour on board the train and simulation of improper behaviour. A leak of video clips simulating, for example, a socially inappropriate action (such as a fight or the damage to a carriage) could potentially damage the reputation of the people filmed, if it is not clearly indicated that it is a simulation.

Fairness with harmful bias managed	What is the business consequence level if the AI system is not fair?	Negligible	In this context (people counting, object detection and classification), fairness is not an issue.
Accountability and Transparency	What is the business consequence level if the AI system is not transparent?	Negligible	The new functionalities offered by our models are not accountable nor transparent for the user. On the other hand, as model output has no legal consequences, or is overseen by public transport personnel, there are no consequences at the business level.
Validity and reliability	What is the business consequence level if the AI system is not accurate?	Very High	Accuracy is crucial for the effectiveness; inaccuracies could lead to loss of trust.
	What is the business consequence level if the AI system is not robust?	Very High	Robustness is essential for the system to handle various environmental and ensure reliable performance; the absence could lead to loss of trust.

Table 12 - Business consequences for poor trustworthiness characteristics.

3.1.4 Vulnerability Assessment

Following the Threat Analysis phase, the Vulnerability Assessment aimed to evaluate the potential weaknesses associated with the identified threats affecting the assets of the SUT. Considering the need to prepare the DPIA with the transportation company, initial discussions were carried out and identified as mainly related to privacy preservation, reliability and robustness. These discussions turned to be in line with the results obtained with TrustGuard, as reported in Section 44. The main vulnerabilities of the PassengerCounting asset are related to *Failures or malfunctions of the AI asset*, which can be very high risk because the model's performance is frequently degraded by real-world environmental factors such as poor lighting, fixed camera angles, and occlusions caused by high crowd density in bus environments. While the overall criticality of the system is classified as medium, the business consequences of a privacy breach are rated as high due to the continuous acquisition of personal video data via CCTV, which requires strict adherence to GDPR regulations to avoid significant reputational and regulatory impacts. Furthermore, the system lacks automated mechanisms for detecting model

drift or identifying adversarial inputs, meaning it relies heavily on periodic manual evaluations rather than real-time technical safeguards. More precise results as obtained by TrustGuard are reported in Section 3.1.3. as a summary of Annex I and Annex II.

3.1.5 Risk Analysis

Based on the impact and vulnerability analyses coupled with the severity assessment reported in Table 9 and Table 11, the risk analysis phase is aimed at identifying specific risk levels within the transportation domain. Through detailed consultations with the transportation company, we estimated threat frequencies, system vulnerabilities, and potential impacts. These factors allow for a comprehensive risk assessment across several trustworthiness dimensions, specifically:

- **Threat Level:** the expected frequency at which a specific threat occurs.
- **Vulnerability Level:** the system's susceptibility to threats based on existing control measures.
- **Consequence Level:** the potential impact of a threat on trustworthiness dimensions and the resulting business consequences of such a compromise.
- **Team Maturity:** the Trustworthy AI proficiency of the development teams and the reliability of AI participants."

The corresponding values have been obtained more precisely by using TrustGuard and are mainly associated with the detected vulnerabilities, so mainly related to potential failures and reliability issues, as well as privacy preservation issues. Section 4.2 illustrates more in detail the evaluation done.

3.1.6 Countermeasures

Building on the vulnerabilities and impacts identified in the Transportation LSP, this section introduces the overall approach adopted to mitigate the main risks affecting the SUT. In line with the FAITH AI TAF v1 framework, countermeasures are defined considering the operational characteristics of public transport contexts, where AI components operate in dynamic, real-world environments and support non-safety-critical decision-making processes.

The assessment highlighted that the most relevant risks for the Passenger Counting functionality are related to robustness and reliability under variable environmental conditions, privacy and data protection obligations associated with continuous video acquisition, and limited capabilities for early detection of performance degradation over time. Countermeasures are therefore selected to reduce exposure along these dimensions, prioritising measures that are proportionate to the system's criticality and compatible with pilot deployment constraints.

At a high level, the mitigation strategy combines technical, organisational, and operational measures. Technical actions focus on strengthening data governance practices, supporting stable model behaviour in real operating conditions, and introducing basic monitoring mechanisms to

increase awareness of potential performance degradation. Organisational and procedural measures complement these controls by addressing privacy preservation, access control, and operational accountability in alignment with DPIA requirements agreed with the transportation company.

In parallel, human oversight and adoption aspects are considered essential to ensure appropriate use of system outputs. Clear communication of system limitations and improved interpretability of results are key to avoiding misinterpretation and over-reliance by operators in daily transport operations. Within this process, FAITH tools such as TrustGuard and TrustSense support the structured consolidation of identified controls and their traceability across assessment iterations.

The concrete mitigation status for each identified threat, classified as treated, postponed, or ignored for the Passenger Counting functionality, is presented in detail in Section 3.6.

3.2 Tool Validation Planning & Training Processes

3.2.1 Dry-run

The *Dry-run* activity was carried out through a questionnaire, distributed by ATC, that implemented the first four phases (Cartography, Threat Assessment, Impact Assessment, Vulnerability Assessment) of the FAITH_AI_TAF methodology. At the time of the dry-run the FAITH Golden Table contained vulnerabilities, threats and controls only from ENISA [1]. The evaluation was carried out in April/May 2025 when the AI assets were still in the Design or early Deploying phase and before the bus platform considering the SUT addressing the task of train coach monitoring. Table 13 below shows the outcome of such an activity at the time it was done, so some information could result outdated.

Threat(s)	Vulnerability	Relevant AI Asset for the control question	Template Control Questions	Answer	Comment
Label modification	Use of unreliable sources to **label d**Dry**ata**	Dtr2-LSP2-Transit Residual Detection	*In the context of Dtr2, have you ensured that reliable sources were used to label the data?*	Yes	Labels created by expert engineers with 2-person check
Poisoning	Model vulnerable to data poisoning adversarial attacks	M2-LSP2-Transit Residual Detection	*In the context of M2, have you chosen and defined a*	No	

			resilient model design?		
		M2-LSP2-Transit Residual Detection	*In the context of M2, have you assessed the exposure level of the model to poisoning adversarial attacks?*	No	
Poisoning	Lack of data for increasing robustness to data poisoning attacks	M2-LSP2-Transit Residual Detection	*In the context of M2, have you enlarged the training dataset by adding diverse data to increase its robustness to poisoning attacks?*	Yes	Data augmentation - in progress
Poisoning, Model or data disclosure, Compromise of ML application components, Human error	Poor access rights management	M2-LSP2-Transit Residual Detection	*In the context of M2, have you applied an RBAC model, respecting the least privileged principle?*	No	*not yet, it will be done*
Poisoning	Poor data management	M2-LSP2-Transit Residual Detection	*In the context of M2, have you ensured that the model complies with data security requirements?*	No	
Compromise of ML application components, Poisoning	Undefined indicators of proper functioning, making complex	M2-LSP2-Transit Residual Detection	*In the context of M2, have there been defined indicators of proper functioning of the model?*	No	*not yet, it will be done*

	compromise identification				
Data disclosure, Compromise of ML application components, Lack of cybersecurity awareness, Poisoning	Lack of consideration of attacks to which ML applications could be exposed to	M2-LSP2-Transit Residual Detection	*In the context of M2, have you integrated ML specificities to awareness strategies and ensured all ML stakeholders are receiving it?*	No	
Human error, Poisoning	Use of uncontrolled data	M2-LSP2-Transit Residual Detection	*In the context of M2, have you controlled all data used by the model?*	Yes	Data controlled by expert engineers with 2-person check
Poisoning	Use of unsafe data or models (e.g. with transfer learning)	M2-LSP2-Transit Residual Detection	*Is the base model used for M2 reliable?*	Yes	Pre-trained model pulled directly from Ultralytics
Poisoning	No detection of adversarial samples in the training dataset	Dtr2-LSP2-Transit Residual Detection	*Does Dtr2 include adversarial samples?*	No	
Poisoning, Model or data disclosure, Compromise of ML application components	Weak access protection mechanisms for ML model components	M2-LSP2-Transit Residual Detection	*In the context of M2, have you ensured that the model complies with identity management, authentication, and access control policies?*	Yes	
Model or data disclosure,	Existence of unidentified	M2-LSP2-Transit Residual Detection	*In the context of M2, have you conducted risk*	No	

Compromise of ML application components	disclosure scenarios		*analysis to identify all disclosure scenarios?*		
Model or data disclosure, Compromise of ML application components	Lack of security process to maintain a good security level of the components of the ML application	M2-LSP2-Transit Residual Detection	*Have you implemented processes to maintain security levels of M2 components over time?*	No	
Data disclosure	The model can allow private information to be retrieved	M2-LSP2-Transit Residual Detection	*Does the M2 model respect differential privacy?*	No	
		M2-LSP2-Transit Residual Detection	*Have you designed the M2 model 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 ML algorithm training	M2-LSP2-Transit Residual Detection	*In the context of M2, have you used federated learning to minimise the risk of data breaches?*	No	
Data disclosure, Model disclosure, Compromise of ML application components	Too much information available on the model architecture	M2-LSP2-Transit Residual Detection	*In the context of M2, have you limited the available information about the model architecture?*	No	

Data disclosure, Model disclosure, Compromise of ML application components	Too much information about the model given in its outputs	M2-LSP2-Transit Residual Detection	*In the context of M2, have you reduced the output information given by the model?*	Yes	The model output is displayed only in debug sessions
Compromise of ML application components	Existence of several vulnerabilities because the ML application was not included into process for integrating security into projects	M2-LSP2-Transit Residual Detection	*In the context of M2, have it been ensured that the model has followed the global process for integrating security?*	No	*not yet, it will be done*
Compromise of ML application components	Use of vulnerable components (among the whole supply chain)	M2-LSP2-Transit Residual Detection	*In the context of M2, have you checked the vulnerabilities of the components used so that they have an appropriate security level?*	No	*not yet, it will be done*
Compromise of ML application components	Existence of several vulnerabilities because ML specificities are not integrated to existing policies	M2-LSP2-Transit Residual Detection	In the context of M2, have there been integrated model specifities to existing security policies?	No	
Compromise of ML application components	Existence of several vulnerabilities because ML application does not comply with	M2-LSP2-Transit Residual Detection	*In the context of M2, have you ensured that the model complies with the relevant security policies?*	No	*not yet, it will be done*

	security policies				
Compromise of ML application components	Contract with a low security third party	M2-LSP2-Transit Residual Detection	*Have you ensured that the M2 model complies with third parties' security requirements?*	No	*not yet, it will be done*

Table 13 - Threats and vulnerability of the AI assets related to the Transit Residual Detection (in the design phase) as evaluated during the Dry-run phase.

Threat(s)	Vulnerability (from Golden Table)	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)
Label modification	Use of unreliable sources to **label data**	Dtr1 - LSP2-PassengerCounting	In the context of Dte1, have you ensured that reliable sources were used to label the data?	Yes	
Poisoning	Model vulnerable to data poisoning adversarial attacks	M1- LSP2-PassengerCounting	In the context of M1, have you chosen and defined a resilient model design?	Yes	
		M1- LSP2-PassengerCounting	In the context of M1, have you assessed the exposure level of the model to poisoning adversarial attacks?	No	

Poisoning	Lack of data for increasing robustness to data poisoning attacks	M1- LSP2-PassengerCounting	In the context of M1, have you enlarged the training dataset by adding diverse data to increase its robustness to poisoning attacks?	No	
Poisoning, Model or data disclosure, Compromise of ML application components, Human error	Poor access rights management	Dtr1- LSP2-PassengerCounting	In the context of Dte1, have you applied an RBAC model, respecting the least privileged principle?	No	not yet, it will be done
Poisoning	Poor data management	M1- LSP2-PassengerCounting	In the context of M1, have you ensured that the model complies with data security requirements?	No	not yet, it will be done
Compromise of ML application components, Poisoning	Undefined indicators of proper functioning, making complex compromise identification	M1- LSP2-PassengerCounting	In the context of M1, have there been defined indicators of proper functioning of the model?	Yes	periodic tests with known results
Data disclosure, Compromise of ML application components, Lack of cybersecurity	Lack of consideration of attacks to which ML applications could be exposed to	M1- LSP2-PassengerCounting	*In the context of M1, have you integrated ML specifities to awareness strategies and ensured all ML stakeholders*	No	*not yet, it will be done*

awareness, Poisoning			*are receiving it?*		
Human error, Poisoning	Use of uncontrolled data	M1- LSP2-PassengerCounting	*In the context of M1, have you controlled all data used by the model?*	Yes	
Poisoning	Use of unsafe data or models (e.g. with transfer learning)	Dtr1- LSP2-PassengerCounting	*Is Dte1 reliable?*	Yes	
Poisoning	Lack of control for poisoning	M1- LSP2-PassengerCounting	*Have you integrated poisoning control after the "model evaluation" phase of M1?*	Yes	we are developing diagnostics on the images acquired by the cameras
Poisoning, Model or data disclosure, Compromise of ML application components	Weak access protection mechanisms for ML model components	M1- LSP2-PassengerCounting	*In the context of M1, have you ensured that the model complies with with identity management, authentication, and access control policies ?*	Yes	
Model or data disclosure, Compromise of ML application components	Existence of unidentified disclosure scenarios	M1- LSP2-PassengerCounting	*In the context of M1, have you conducted risk analysis to identify all disclosure scenarios?*	Yes	

Model or data disclosure, Compromise of ML application components	Lack of security process to maintain a good security level of the components of the ML application	M1- LSP2-PassengerCounting	*Have you ensured that M1 complies with protection policies and that these are integrated to security operations processes?*	Yes	
		M1- LSP2-PassengerCounting	*Have you ensured that appropriate protections are deployed for the test environment of M1?*	Yes	
Data disclosure	The model can allow private information to be retrieved	M1- LSP2-PassengerCounting	*Does the M1 model respect differential privacy?*	Yes	
		M1- LSP2-PassengerCounting	*Have you designed the M1 model in such a way that will provide only the necessary information and will not disclose any private information?*	Yes	
Data disclosure,	Disclosure of sensitive data for ML algorithm training	M1- LSP2-PassengerCounting	*In the context of M1, have you used ferated learning to minimise the risk of data breaches?*	No	

Data disclosure, Model disclosure, Compromise of ML application components	Too much information available on the model architecture	M1- LSP2-PassengerCounting	*In the context of M1, have you limited the available information about the model architecture?*	Yes	
Data disclosure, Model disclosure, Compromise of ML application components	Too much information about the model given in its outputs	M1- LSP2-PassengerCounting	*In the context of M1, have the output information given by the model been reduced?*	Yes	The model output are displayed only in debug sessions
Compromise of ML application components	Existence of several vulnerabilities because the ML application was not included into process for integrating security into projects	M1- LSP2-PassengerCounting	*In the context of M1, have it been ensured that the model has followed the global process for integrating security?*	Yes	
Compromise of ML application components	Use of vulnerable components (among the whole supply chain)	M1- LSP2-PassengerCounting	*In the context of M1, have you checked the vulnerabilities of the components used so that they have an appropriate security level?*	Yes	
Compromise of ML application components	Existence of several vulnerabilities because ML specificities	M1- LSP2-PassengerCounting	*In the context of M1, have there been integrated model*	Yes	

	are not integrated to existing policies		*specifities to existing security policies?*		
Compromise of ML application components	Existence of several vulnerabilities because ML application do not comply with security policies	M1- LSP2-PassengerCounting	*In the context of M1, have you ensured that the model complies with the relevant security policies?*	Yes	
		M1- LSP2-PassengerCounting	*Have you included the M1 model into the asset management processes?*	Yes	
Compromise of ML application components	Contract with a low security third party	M1- LSP2-PassengerCounting	*Have you ensured that the M1 model complies with third parties' security requirements?*	Yes	

Table 14. Threats and vulnerability of the AI assets related to the Passenger Counting (in the develop and deploy phase) as evaluated during the Dry-run phase.

3.2.2 Plan development

During the Oslo plenary meeting held on January 2026, a detailed Implementation Plan for the Transportation LSP has been developed. The temporal window considered is the year 2026 from January (M25 - Month 25th) to December (M36 - Month 36th). The following table reports the updated version of this plan in April 2026 (M28 - Month 28th). The LSP Risk Assessor has been nominated and indicated to the rest of the Consortium.

Task	Assigned to	Progress	Start	End
Attend training for AI Model Passport (pythonic library)	CNR / MER MEC	100%	M25	M26
Attend training for AI Data Passport	CNR / MER MEC	100%	M25	M26
Attend training for trustSENSE	CNR / MER MEC	100%	M23	M23
Attend training for trustGUARD	CNR / MER MEC	100%	M26	M27

Attend training for SSM	CNR / MER MEC	0%	M32	M33
Attend introduction to FAITH STM Platform registration and workflows	CNR / MER MEC	0%	M33	M33

Table 15 - Training and support activities.

Task	Assigned to	Progress	Start	End
Deploy SUT in SANDBOX	CNR / MER MEC	100%	M20	M24
Deploy SUT in REAL ENVIRONMENT	CNR/ MER MEC	20%	M25	M38
Use TrustSense for team maturity assessment (early deployment period)	Risk Assessor + End Users	66%[11]	M25	M30
Provide formal feedback on Usability and Usefulness of TrustSense	Risk Assessor	66%	M26	M30
Implement TrustSense mitigations	CNR / MER MEC / End user	10%	M27	M51
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%	M26	M28
Iterate AI Model Hub v2 based on LSP feedback	FORTH / CNR	0%	M28	M32
Use TrustGuard v1 for in-isolation AI Model Asset assessment @ deployment phase	Risk Assessor	100%	M26	M28
Use TrustGuard v1 for in-isolation (Input) Data Asset assessment @ deployment phase	Risk Assessor	100%	M26	M28
Provide formal standardised feedback on Usability and Usefulness of TrustGuard v1 during deployment phase	Risk Assessor	100%	M26	M28
Apply 1 or more Controls to AI Model Asset or Data Asset during deployment phase based on TrustGuard recommendations	CNR / MER MEC	0%	M29	M33
Update and re-run in-isolation AI Model Asset and/or Data Asset assessment	Risk Assessor	0%	M33	M36

[11] Only CNR and MERMEC teams used TrustSense tool. Domain Expert will do it in the next month

including new Control(s) @ deployment phase using TrustGuard v1				
On-going integration of TrustGuard v2 based on LSP feedback @ all 3 lifecycle phases (inc. on-going Knowledgebase updates)	ATC	0%	M30	M36
Use FAITH STM Platform	Risk Assessor	0%	M34	M36
Provide formal standardised feedback on Usability and Usefulness of FAITH STM Platform	Risk Assessor	0%	M36	M36
Iterate FAITH STM Platform	UOS		M34	M36
Create SSM system model	CNR / MER MEC / UOS	0%	M34	M34
Use SSM for in-context SUT assessment	Risk Assessor	0%	M34	M35
Provide formal standardised feedback on Usability and Usefulness of SSM	Risk Assessor	0%	M35	M35
On-going interaction of SSM v2 based on LSP feedback (inc. on-going Knowledgebase updates)	UOS	0%	M36	M45
Deploy SUT in Replication Pilot	CNR / MER MEC	0%	M36	M38

Table 16 - FAITH tools implementation and iteration 2026.

3.2.3 Tool Training and validation

The implementation and operational use of the FAITH toolset, specifically TrustSense, TrustGuard, and the Data and AI Model Passports, have been supported by a series of dedicated training and collaborative sessions. These meetings were instrumental in ensuring that the LSP2 stakeholders were acquainted with the tools and able to evaluate the trustworthiness and security of the Transportation LSP SUTs.

TrustSense Implementation and Preliminary Use. The initial training and application of the TrustSense tool took place on November 10th 2025. After that training, the tool was utilized to gather results about the maturity level of the stakeholders involved, namely the AI maturity of the MERMEC, CNR and end-users' teams. Results have been reported as screenshots. Current records suggest that the primary hands-on execution at this stage was handle by each team Assessor, with results disseminated to the broader group for review.

Data and AI Model Passports Training. A comprehensive session dedicated to the Data and AI Model Passports was held on February 6th 2026. This meeting was led by the FORTH team and served as a critical knowledge-transfer event for the project. The session saw broad participation

from across the MERMEC, and CNR teams. This collaborative approach ensured that both the developers of the models and the data managers were aligned on the transparency and documentation standards required by the FAITH framework.

TrustGuard Security Assessment. The training of the TrustGuard vulnerability assessment were conducted on February 27th 2026. This session focused on elucidating how the tool can be used. The meeting was attended by the core technical and security team from MERMEC and CNR, particularly the Risk Assessor for LSP2. After that, the Risk Assessor used the tool to evaluate the system, identify potential security threats and receive the mitigation recommendations.

After the use of each tool, feedback was provided responding to the corresponding questionnaire.

3.3 FAITH Tool Validation Activities

3.3.1 TrustSense

The TrustSense tool has been used by the AI developers from both CNR and MERMEC Engineering to assess the trustworthiness level of the team.

The results indicate a trustworthiness level ranging from moderate to high. Naturally, these values should be considered a preliminary reliability assessment, as they will be reassessed throughout the project to monitor and report any improvements resulting from the implementation of recommended risk mitigation strategies.

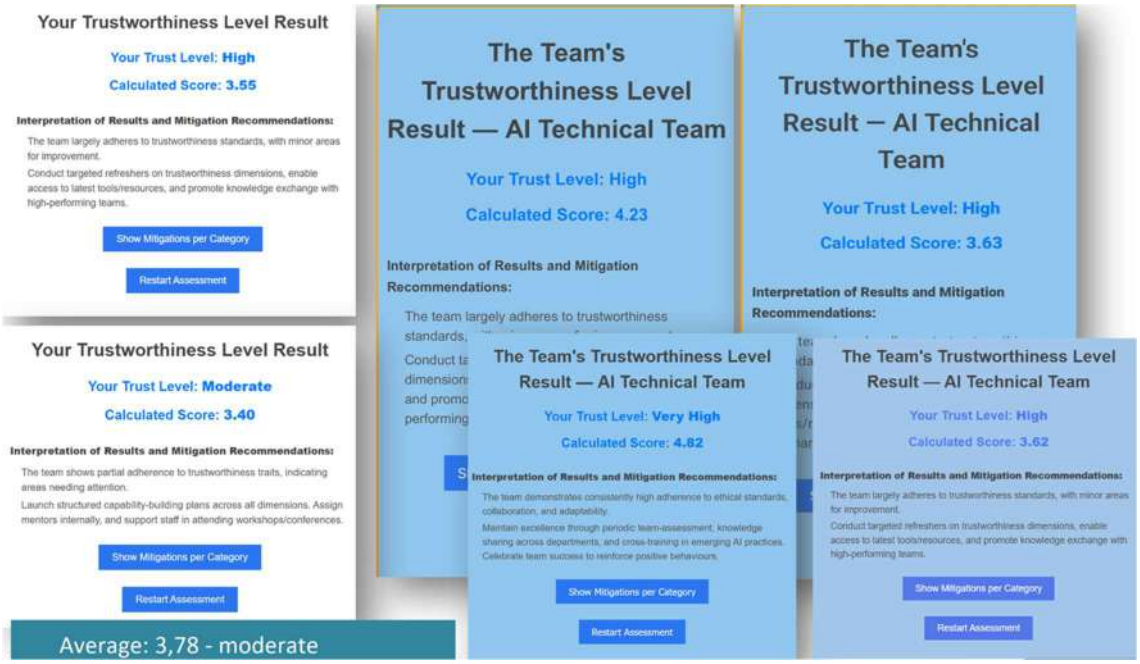

Figure 23 - TrustSense: samples of results collected after the use of the tool.

The TrustSense tool has been used also by one of the end-user team (the bus company in Bologna) to assess the trustworthiness level of the team. The results show a high trust level, see Figure 24.



Figure 24 - TrustSense: results collected in a Domain AI User team (bus company in Italy).

The mitigation actions recommended by the tool, associated with low scores, are related to "*problem solving*" and "*motivation and commitment*". Some examples of mitigation actions are reported in Figure 25 below.

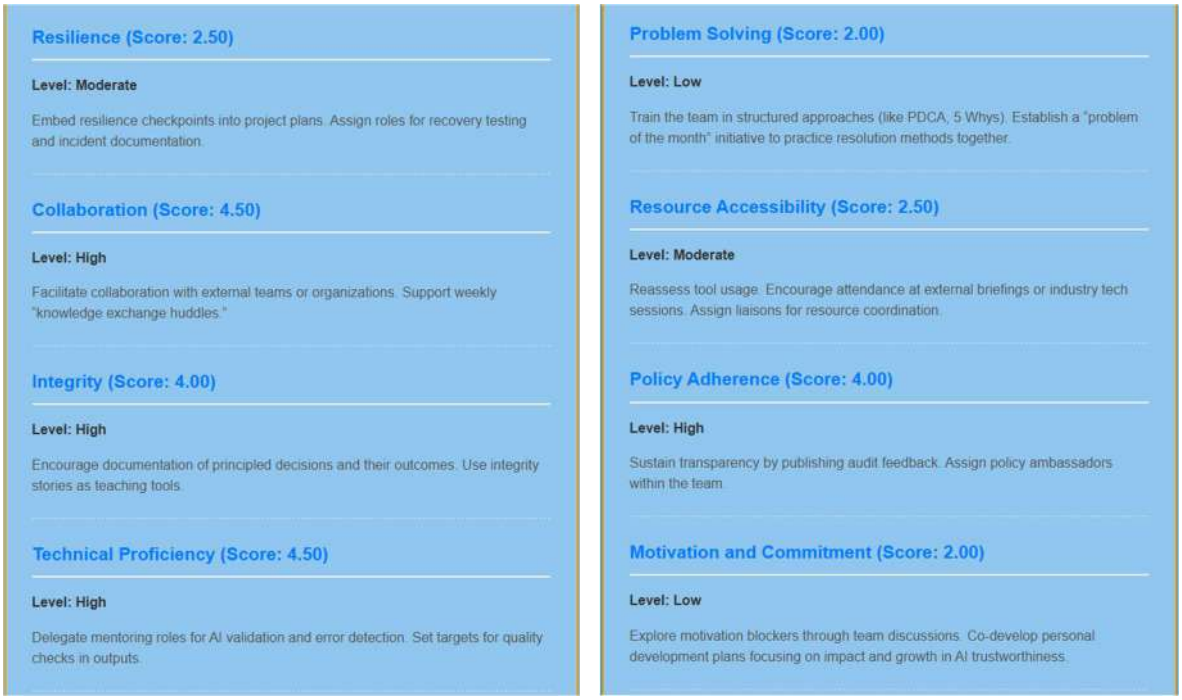

Figure 25 - TrustSense: some examples of recommended mitigation actions.

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

The following sections summarise the activities done to use the Data and AI Model Passport for the SUT assets.

3.3.2.1 Passenger Counting - Data Passport

The Data Passport for the Passenger Counting dataset was created in the FAITH Model Hub. The instance has the name FAITH-LSP2-PassengerCounting (see Figure 26).

The dataset contains still images extracted from onboard CCTV cameras inside buses, collected during regular service. Each image is labelled with the ground-truth number of passengers to train and evaluate computer-vision models for passenger counting.

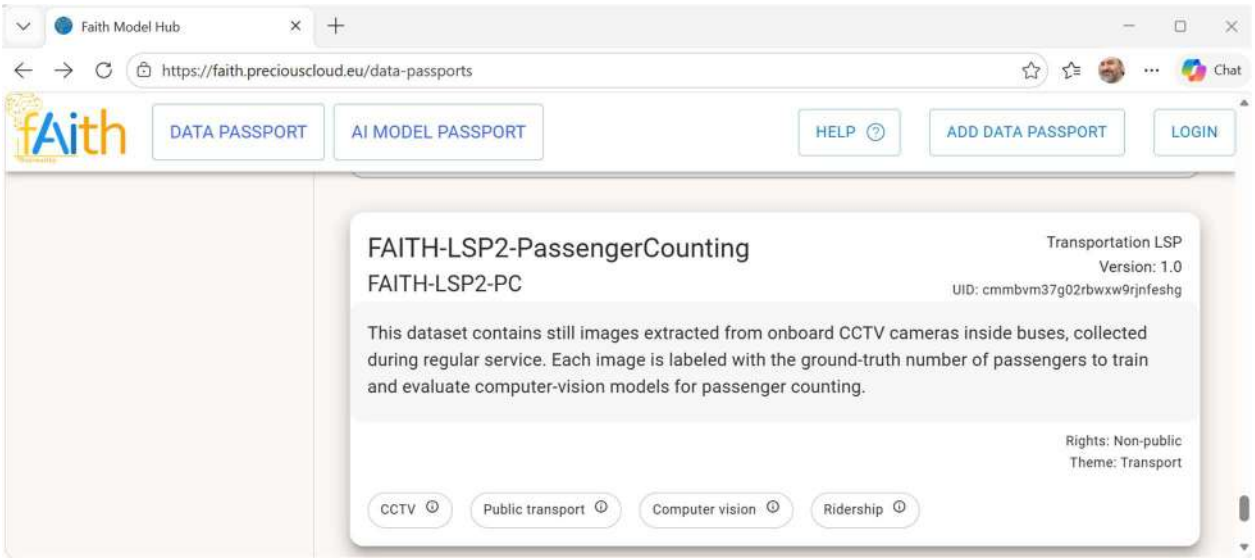

Figure 26 - The Data Passport for the Passenger Counting on the FAITH AI Model Hub.

3.3.2.2 Transit Residual Detection – Data Passport

The Data Passport for the Transit Residual Detection was created in the FAITH AI Model Hub. The instance has the name LSP2-Transit Residual Detection Dataset (see Figure 27).

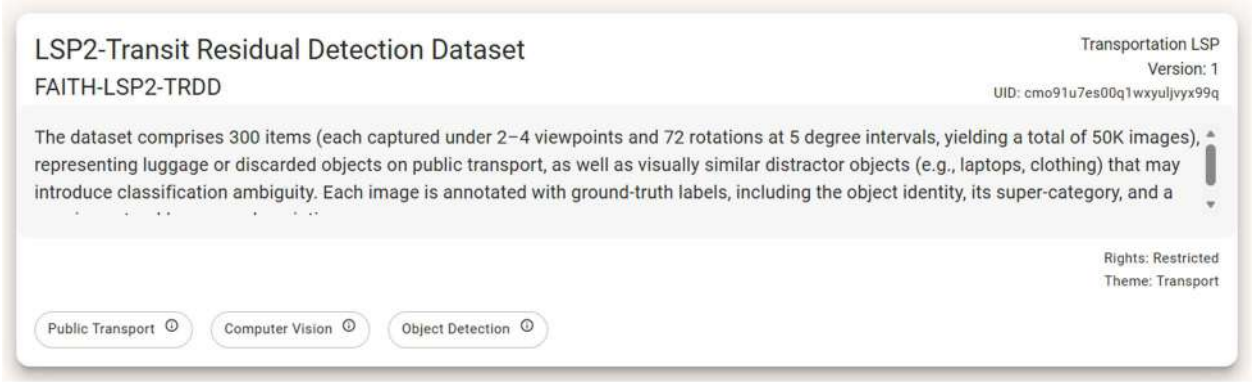

Figure 27 - The Data Passport for the Transit Residual Detection on the FAITH AI Model Hub.

The dataset comprises 300 items (each captured under 2–4 viewpoints and 72 rotations at 5° intervals, yielding a total of 50,000 images), representing luggage or discarded objects on public transport, as well as visually similar distractor objects (e.g., laptops, clothing) that may introduce classification ambiguity. Each image is annotated with ground-truth labels, including the object's identity, its super-category, and a concise natural-language description.

3.3.2.3 AI Model passports

The Python library by FORTH to automate the creation of the AI model passports was available in February 2026. Despite the short timing, it was successfully used within our code. As evidence, the screenshots in Figure 28, Figure 29, Figure 30, Figure 31, and Figure 32 show the full execution of the pipeline steps needed to train the Passenger Counting Model (it is the same for the Transit Residual Detection). This processing was done with a small subset of images to demonstrate the successful integration of the software. The result can be found in MLFlow with the name LSP2_passenger_counter.

Unfortunately, the automatic upload of the data record in the FAITH AI Model Hub did not succeeds, and this issue was reported to the FORTH developer team.

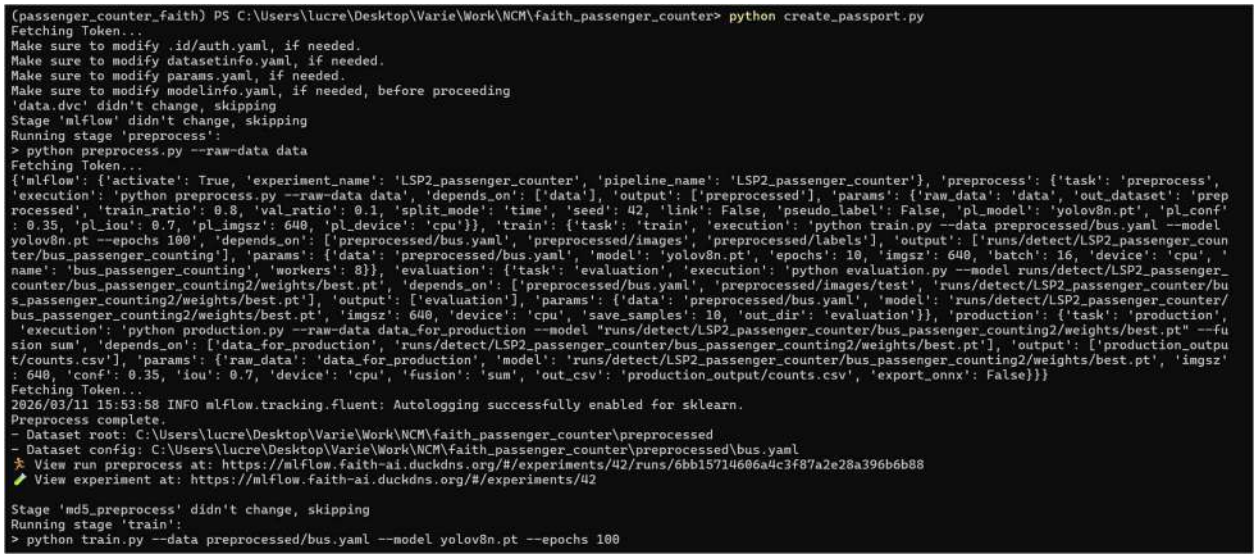

Figure 28 – AI Passport: execution of the "preprocess" stage of the pipeline.

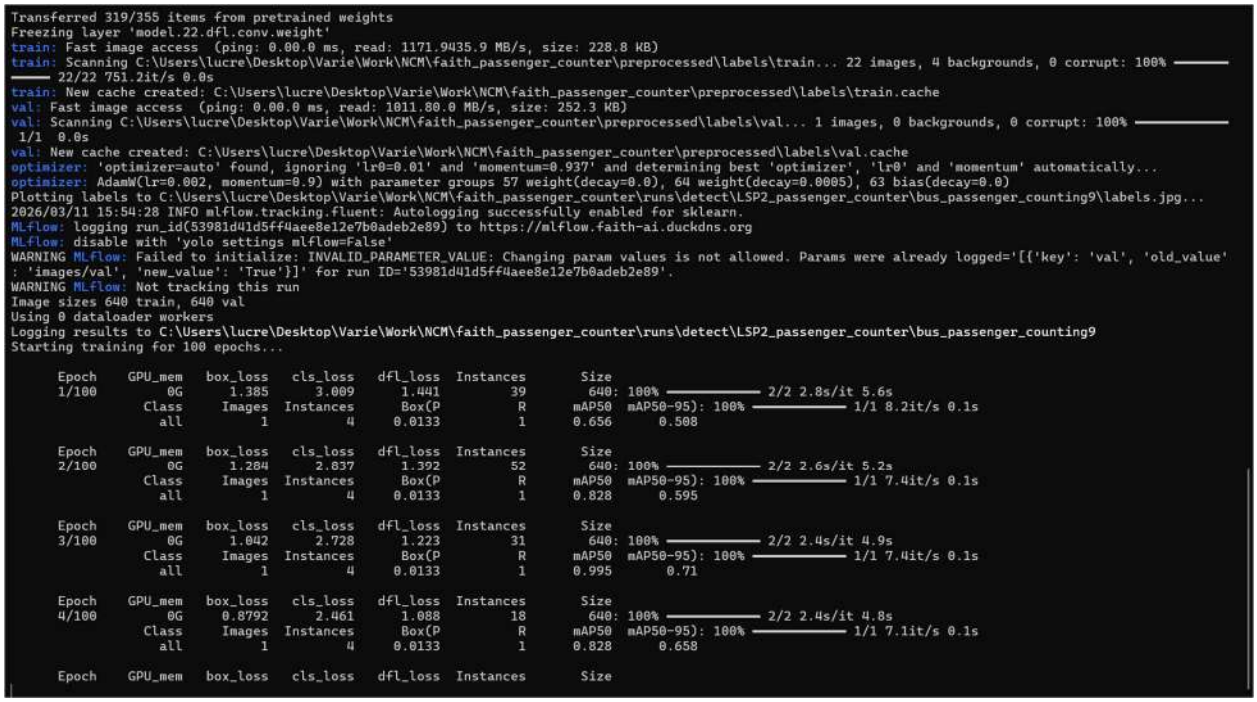

Figure 29 - AI Passport: execution of the "train" stage.

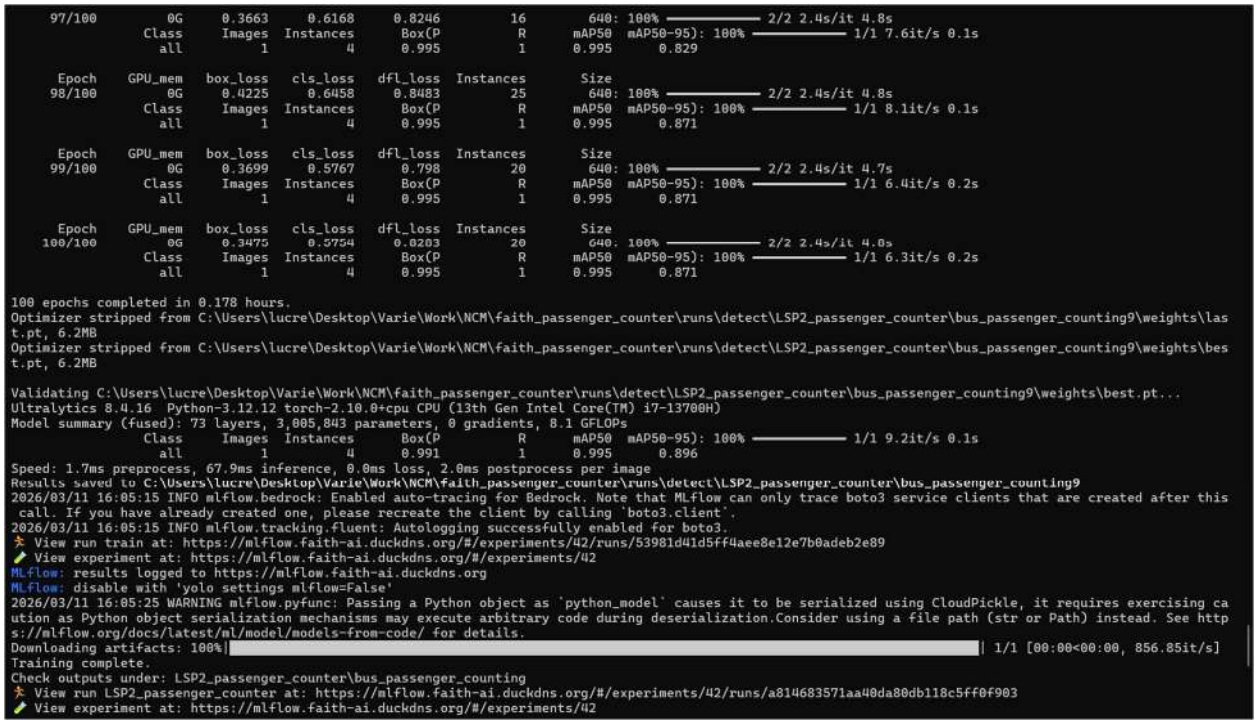

Figure 30 - AI Passport: the training is completed.

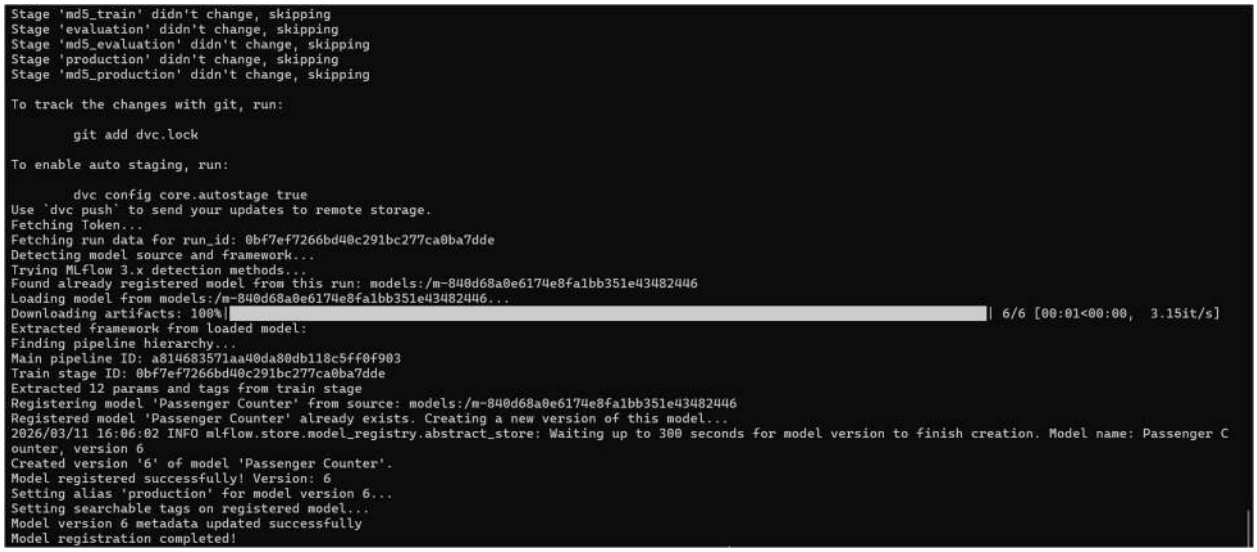

Figure 31 – AI Passport: the model has been registered successfully on MLFlow.

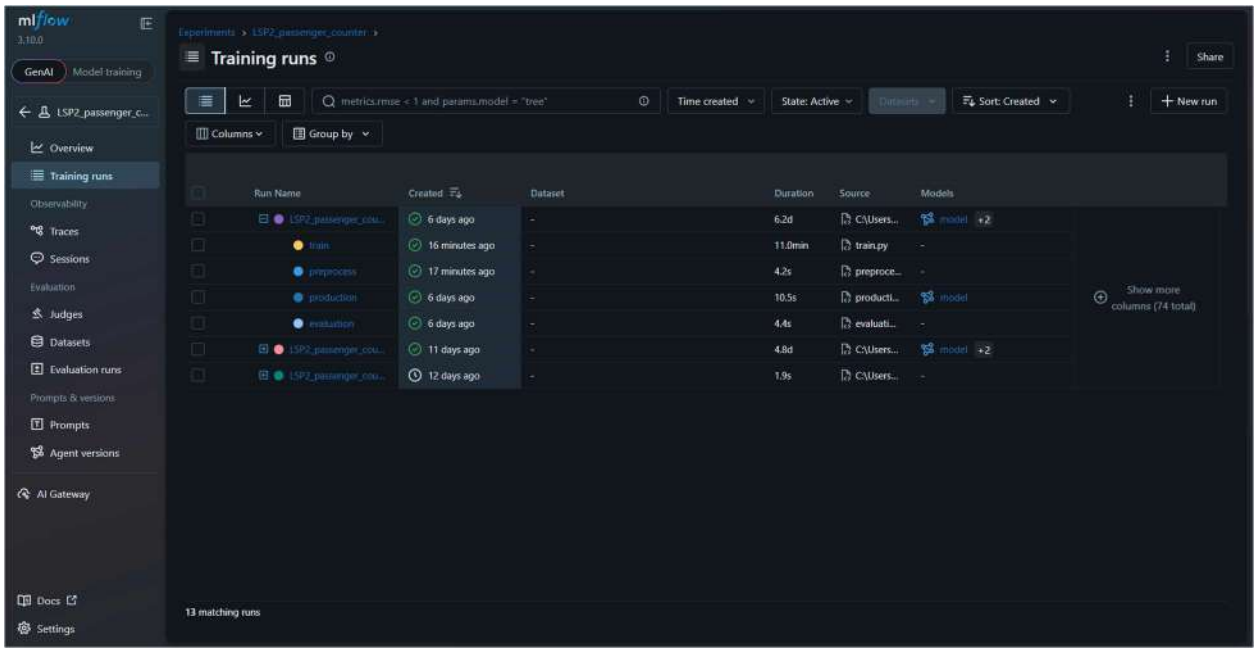

Figure 32 - AI Passport: The LSP2_passenger_counter runs can be checked on MLflow.

3.3.3 TrustGuard

The TrustGuard tool is a core component of the FAITH framework, designed to support structured threat analysis, risk assessment, and the identification of mitigation strategies for AI-based systems.

Within the Transportation LSP, TrustGuard is being progressively adopted to assess risks associated with the SUT, focusing on asset-level analysis in alignment with the FAITH TAF V1

methodology. Its application is also aligned with internationally recognized AI risk management frameworks, including ISO/IEC 23894:2023 as well as with the emerging regulatory requirements defined by the EU Artificial Intelligence Act.

At the current stage of the project, TrustGuard is in an initial validation phase and is being used by technical teams from CNR and MERMEC Engineering and the nominated Risk Assessor to support the following activities:

- **Structured identification of threats**, affecting AI models, datasets, system infrastructure, and human actors, in line with the "MAP" function of the NIST AI RMF;
- **Assessment of vulnerabilities**, including both technical vulnerabilities (e.g., model robustness, sensor reliability, data quality) and organizational aspects (e.g., user training, operational procedures), consistent with ISO/IEC 23894 guidelines;
- **Risk evaluation**, based on the estimation of likelihood and impact, supporting the classification of risks in accordance with the risk-based approach promoted by the EU AI Act;
- **Definition of mitigation measures and controls,** addressing key trustworthiness dimensions such as robustness, transparency, privacy, and human oversight.

Currently, TrustGuard has been used to evaluate the data and AI model for the Passenger Counting functionality and the data asset for the Transit Residual functionality. The results are summarized as follows

The classification of the criticality level of the SUT assets are consistent with the analysis reported in Section 3.1.1, having the Passenger Counting classified at a "Moderate" criticality level (see Annex I) and the Data asset for the Transit Residual Detection classified at a "Low" criticality level (see Annex II).

3.4 Threat Analysis

The results of the threat analysis are reported below for the evaluated assets.

3.4.1 Passenger Counting – Dataset and AI model

The threat assessment for the LSP2-PassengerCounting asset evaluates the historical frequency and perceived risk of various AI-specific and operational threats. The assessment indicates a generally stable environment with low threat levels across most categories. Hereby we report a summary of the results:

High-Frequency and Contextual Threats

These threats occur relatively often or are inherent to the specific operational environment of a passenger-counting system using CCTV.

- Dynamic User Behaviour Impacts: High to Medium Frequency. Variations in passenger movement and behaviour patterns significantly impact detection and counting performance in real-world conditions.
- Personal Video Data Acquisition: occurs continuously. This is a constant contextual threat as the system relies on CCTV cameras to capture video of passengers, requiring continuous management under GDPR.
- Misclassification of Objects: Medium Frequency. Challenging conditions like crowd density, occlusions, and varying lighting lead to frequent incorrect detections or classifications.
- Failure or Malfunction of ML Application: Once a Year. Identified as a recurring operational risk.
- Human Error: Once a Year. Identified as a regular occurring threat within the domain.

Low-Frequency Threats

These threats are noted as having a theoretical risk but are rare in this specific application.

- Evasion: Once every 10 years. Intentionally crafted inputs to cause wrong predictions are considered very low risk.
- Reweighing Inaccuracies: Low Frequency. Not a critical component for single-class object detection, though minor issues can arise from imbalanced training data.
- Inspection Neglect due to Biased Path Sampling: Low Frequency. While the system doesn't perform path sampling, limited camera coverage and blind spots can lead to partial observation.
- Poisoning: Low Frequency/Treated. While listed generally as low, it is specifically addressed through data integrity checks prior to deployment.

Never Occurring / Not Applicable Threats

The vast majority of technical AI threats are marked as "**Never**" or "**N/A**" because the system does not process user-provided text or perform autonomous navigation.

- Adversarial Attacks: Oracle attacks, and use of adversarial examples (white/grey box conditions).
- Data & Model Disclosure: model or data disclosure, label modification, and privacy leakage via footage (noted as "Never" despite being a GDPR risk).
- Generative AI Specifics: prompt attacks, propagation of misinformation, and RAG-related failures (Not applicable as the system does not use a Large Language Model).
- Infrastructural/Regulatory: Compromise of ML components, regulatory non-compliance, and cybersecurity incidents not reported.

- Fairness & Bias: Dataset bias from water imagery, bias in learning recommendations, and unfair discrimination in outputs.
- Operational/Navigation: Unsafe routes near sensitive structures, tampered mission parameters, and suboptimal route execution under uncertain currents.

Domain-Specific Exceptions

The assessment highlights that certain common AI threats are not applicable to this specific asset:

Autonomous & Text-Based Threats: threats like "Prompt attacks" or "Misinformation" are irrelevant because the system does not process user-provided text inputs or use autonomous agents.

Pathing & Navigation: threats related to "Unsafe Route Execution" are marked as "Never" because the system operates on fixed camera inputs and does not perform navigation.

Overall Risk Summary

Criticality Assessment: Medium.

General Risk Score: most individual threats are scored as “Low”, with the exception of Evasion, which was tagged as “High” in the results section but marked as "Once every 10 years" in the frequency section, suggesting a high impact but very low probability.

Primary Mitigation Strategy: the report largely relies on the "relative stability of the operational environment" and periodic manual evaluations rather than automated real-time drift or fairness controls.

Business Consequence Assessment

The report links these threats to potential business impacts, concluding that while most risks are low, the consequences of a breach are notable:

Safety/Security Compromise: rated as “Moderate to High”. A compromise could lead to service disruption and potential data breaches involving sensitive CCTV data, leading to operational and reputational damage.

Interpretability & Transparency: the consequence level is rated as “Low”, as the system's lack of complex text output makes interpretability less critical for business continuity than security or accuracy.

3.4.2 Transit Residual Detection - Dataset

The threat assessment for the LSP2-TransitResidualDetection AI asset evaluates the historical frequency and perceived risk of various AI-specific and operational threats. The assessment indicates *a generally stable environment with low threat levels across most categories.*

The summary of the findings is as follows:

High-Frequency and Contextual Threats

Only a few threats were identified as occurring with any notable frequency. These are largely driven by the garbage/residual object detection task and is specific to the real-world operational environment:

Use of adversarial examples crafted under white, or grey-box conditions, Labelling, and Poisoning: identified as occurring "once every year". Even if this risk is infrequent, the availability of the video stream and the training dataset may enable the generation of adversarial examples that degrade the ability of the methods to identify unattended items (e.g., adversarial colour patterns).

Failure or malfunction of ML applications is a frequent issue in such complex tasks and can occur with very high frequency (much higher than twice a year), due to differences in lighting conditions, shadows, patterns, and variability in items. Even if the consequences of such detection failures are relatively low, the high frequency makes this a major drawback of deep learning-based video recognition methods. This threat is strongly related with "Regulatory and legal non-compliance/Personal video data acquisition" which can occur whenever the failure of the system leads to the acquisition of non-anonymized frames of faces of the passengers.

Model drift leading to inaccurate assessments, Biased training data, Limited diversity in training data, Lack of critical information in training and testing data, misclassification of structural defects and objects, and training data quality issues due to acquisition procedures are highly correlated problems. The variability among luggage, garbage, and abandoned items makes it difficult to obtain accurate statements, and performance tends to degrade over time due to a progressive shift in data distribution.

Over-reliance on model predictions: estimated as occurring once every two years, though in practice it may be more frequent due to the limited significance of the selected time scale for such a large application domain (in terms of the number of cameras and video acquisitions). If predictions are used to identify buses requiring cleaning and/or assess risks related to abandoned luggage, this may lead to inappropriate deployment of security and cleaning personnel. Extensive use of these AI models could also reduce the current level of attention and vigilance of such critical personnel, potentially leading to issues in cases where security risks are incorrectly assessed or not detected by the model.

Lack of robustness checks is an uncommon threat (once every two years), although it can occur whenever the model is not sufficiently tested and retrained to account for drift in the data distribution, which may arise over short temporal windows due to changes in fashion trends. This is strongly related with Changes in external conditions, once every 5-year threat.

Unfair discrimination in AI outputs, occurring approximately once every five years. Particular items (e.g., ethnic, rare, or underrepresented cases) may be misclassified as anomalies due to

their deviation from the training domain, leading to infrequent but potentially unfair discrimination. This is strongly related to insufficient safeguards against bias, which may fail to prevent discriminatory behaviour in complex environments such as the transportation domain, particularly with respect to smaller or underrepresented groups.

Low-Frequency and "Never" Occurring Threats

Most identified threats were marked as "Never" having occurred within this domain. This suggests a highly controlled or limited-exposure environment.

Domain-Specific Exceptions

The assessment highlights that certain common AI threats are not applicable to this specific asset:

Autonomous & Text-Based Threats: threats like "Prompt attacks" or "Misinformation" are irrelevant because the system does not process user-provided text inputs or use autonomous agents.

Pathing & Navigation: threats related to "Unsafe Route Execution" are marked as "Never" because the system operates on fixed camera inputs and does not perform navigation.

3.5 Initial Risk Assessment results

3.5.1 Passenger Counting – Dataset and AI model

TrustGuard evaluated the system across the seven distinct dimensions of AI trustworthiness, analysing specific threats, risk levels, and the implementation status of various mitigation controls.

Accountable and Transparent

The assessment for accountability focuses on the visibility of system processes and the prevention of unauthorized access or disclosure.

Risk Profile: all identified risks, including Model or data disclosure, Compromise of ML application components, and Human error, are categorized as Low risk.

Assessment results: the system met all foundational requirements, including the systematic conduct of bias audits, the tracking of data provenance for traceability, and the engagement of legal counsel for data source review. Traceability is ensured via the use of the Data and AI Model Passport, which will be shared with the transportation companies in the replication phase.

Actions: because the foundational architecture is deemed robust, no additional mitigation actions were required at this stage.

Fair with Harmful Bias Managed

This property examines how the system handles potential biases and protects against poisoning or regulatory non-compliance.

Risk Profile: all risks under this category are rated as Low risk.

Proposed Controls: automated fairness testing tools, embedding invariance testing, and adaptive online learning.

Ignored Controls: several advanced mitigations, such as adaptive online learning and automated fairness testing tools, were marked as Ignored. The motivation is reported below.

Motivation: the assessors noted that automated fairness testing is unnecessary because the system performs aggregate object detection and counting rather than individual-level decision-making. Furthermore, the operational environment is relatively stable, meaning model updates can be performed in a controlled, manual manner rather than requiring constant automated re-training.

Actions: no actions required at this stage. However, the evaluation during the replication pilot will further check whether system failures correspond to specific biased conditions (e.g., ethnicity of passengers).

Safe

The safety assessment prioritizes the prevention of system manipulation that could lead to incorrect outputs.

Risk Profile: standard threat risks, including Evasion, are rated as Low risk. This is because we consider Evasion as mainly an intentional threat.

Proposed Controls: adversarial training, adversarial example detection, and post-evaluation poisoning control

Assessment Status: poisoning is marked as treated as motivated in the following paragraph, while the adversarial detection will be better addressed when considering ML malfunctioning due to environmental changes.

Motivation: poisoning control is considered treated because rigorous data integrity and validation checks are performed prior to deployment. Other adversarial controls were ignored as the risk of intentional evasion is low in this specific transit context.

Secure and Resilient

This dimension focuses on the system's ability to withstand attacks and maintain functionality under stress.

Risk Profile: most threats are Low risk, though Evasion for the FAITH-specific asset is rated as High.

Proposed Controls: input modification, use of non-transferable models, and output information reduction.

Assessment Status: poisoning control has been treated, as motivated also above, while input modification is postponed (i.e., planned for the next period).

Motivation: input modification is planned but delayed to align with future development cycles. Output reduction was ignored because the system design inherently protects the model by only providing aggregated counts, never exposing internal parameters

Actions: apply input modification techniques to ensure more robust results.

Explainable and Interpretable

This category ensures that the system's logic and outputs can be understood by human operators.

Risk Profile: threats like Propagation of misinformation and Reduced operator accountability are rated as Low risk.

Proposed Controls: retrieval-augmented generation (RAG), entropy-based measures, and token-level probability estimation

Assessment status: proposed controls are ignored based on the motivation below.

Motivation: the postposed controls are strictly not applicable to this system. They are designed for Large Language Models (LLMs), whereas this asset is a vision-based counting tool that does not process or generate text.

Actions: no actions.

Privacy Enhanced

The assessment confirms the protection of sensitive visual data throughout its lifecycle.

Risk Profile: all privacy-related threats, including Personal video data acquisition and Privacy leakage via inspection footage, are rated as Low risk.

Proposed Controls: encryption, role-based access control, and poisoning control.

Assessment status: the proposed controlled are treated for the motivation below.

Motivation: the system uses encryption mechanisms, role-based access controls, and redaction of sensitive visual elements before storage or sharing. Moreover, privacy is maintained by the system's fundamental architecture: raw video is processed into anonymized aggregate data, ensuring no sensitive personal information is stored or transmitted to the control station.

Actions: no actions, unless strengthening the security levels.

Valid and Reliable

This property evaluates whether the system performs as intended consistently and handles external changes effectively.

Risk Profile: All risks are Low.

Proposed Controls: drift control services, automated malfunction detection, and prompt-attack fine-tuning.

Assessment status: drift control and malfunction assessment tools are postponed (i.e., planned for the next release), based on the countermeasures proposed.

Motivation: drift and malfunction tools are planned for future adoption to improve long-term reliability. Prompt-attack mitigations were ignored because the system has no text-input interface, and path re-planning is irrelevant as the cameras are fixed.

Actions: test drift and malfunctioning-aware strategies to mainly cope with varying environmental conditions.

3.5.2 Transit Residual Detection - Dataset

TrustGuard evaluated the system across the seven distinct dimensions of AI trustworthiness, analysing specific threats, risk levels, and the implementation status of various mitigation controls.

Accountable and Transparent

The assessment for accountability focuses on the visibility of system processes and the prevention of unauthorized access or disclosure.

Risk Profile: all identified threats, including Model or data disclosure, and Compromise of ML application components are categorized as Low risk.

Assessment status: mitigations focus on conducting a risk analysis of the ML model. This will be carried out after the dataset has been integrated into the enhancement of the corresponding model.

Fair with Harmful Bias Managed

This property examines how the system handles potential biases and protects against poisoning or regulatory non-compliance.

Risk Profile: all threats under this category (including poisoning, Lack of critical information in training and testing data, Biased training data, Uniform sampling bias, Failure to quantify bias, and Inadequate assessment of fairness risks) are rated as Low risk.

Ignored Controls and Comments: the mitigation strategies have either already been implemented or postponed due to the current development stage of the ML model. The recommendation to regularly update the risk register with identified fairness threats has been ignored, as the risk register is not fully applicable to the TrustGuard evaluation conducted on the dataset.

Secure and Resilient

This dimension focuses on the system's ability to withstand attacks and maintain functionality under stress.

Risk Profile: all threats selected are rated Low.

Mitigation control treated unless Reduce the information given by the model which is inherently not applicable as the document clarifies that this is inherently addressed by the system's design: it only provides aggregated results (presence of luggage, and number of marked items) and does not expose images, detailed model information or internal parameters to users.

Privacy Enhanced

The assessment confirms the protection of sensitive visual data throughout its lifecycle.

Risk profile: threats are rated as Low risk, except in the case of personal video data acquisition.

The proposed mitigation strategy assumes that image data are not stored, but only processed, and that only alerts (i.e., model outputs) together with metadata are transmitted to the control station. This has already been clarified in the document and is inherently addressed by the system's design, which provides only aggregated results (e.g., presence of luggage and counts of marked items) and does not expose images, detailed model information, or internal parameters to users.

Valid and Reliable

This property evaluates whether the system performs as intended consistently and handles external changes effectively.

Risk Profile: All risks are Low.

Controls: the mitigation "ensure reliable sources are used" is marked as *Treated*. "Conduct a risk analysis of the ML application" is postponed until after the integration of these data to enhance the underlying model. The mitigation "regularly update environmental models" is ignored, as neither the acquired dataset nor the underlying model makes use of environmental models.

3.6 Countermeasures

Current approaches and mitigations planned (i.e., marked as "*Postponed*" in TrustGuard) are reported in the following sections for the SUT assets.

3.6.1 Passenger Counting

Regarding this functionality, each of the identified threats is associated with one or more mitigation strategies, which have been classified as Treated, Postponed, or Ignored based on their relevance and applicability to the system.

Most of the detected threats are assessed as low risk and, consequently, do not require specific mitigation actions. These include risks related to model or data disclosure, compromise of ML components, human error, regulatory non-compliance, and failures or malfunctions of the ML application. In many cases, the nature of the system—focused on aggregate passenger counting from CCTV footage without individual identification or decision-making—naturally limits the impact of such threats. As a potential future enhancement, the system could incorporate monitoring logic aimed at identifying operating conditions under which the algorithm is at higher risk of failure. In particular, when adverse real-world conditions (e.g., poor lighting, occlusions, camera misalignment, or scene changes) impair reliable detection, the system could flag reduced output confidence, thereby increasing awareness of the limited reliability of the results produced by the affected camera.

Some mitigation strategies have been explicitly ignored as they are not suitable for the context. For instance, fairness-related controls (e.g., automated fairness testing, adversarial debiasing, or embedding invariance checks) are not considered necessary, since the system does not perform individual-level predictions but only aggregate object detection. Similarly, several mitigations designed for text-based or generative AI systems—such as prompt attack defenses, entropy-based measures, or retrieval-augmented generation—have been disregarded because the system does not process or generate textual data. Other controls, such as adaptive learning techniques or continuous retraining strategies, have also been ignored due to the relatively stable operational environment, where periodic and controlled model updates are deemed sufficient.

A limited number of mitigation actions related to adversarial robustness (e.g., adversarial training and detection of adversarial inputs) have been postponed due to the fact that they are currently considered of low practical relevance in the deployment context and are therefore deferred to future development phases.

A small subset of mitigations has been actively treated. Notably, input modification techniques have been adopted to improve robustness against environmental changes, and initial drift control mechanisms have been integrated to support continued model reliability over time. In particular, the planned activities foresee the introduction of periodic, system-level drift monitoring and validation checks, aimed at detecting performance degradation and triggering timely mitigation actions during operation. Additionally, some risks, such as poisoning, are

considered already addressed through existing validation and data integrity procedures applied after model evaluation.

Overall, the assessment indicates that many threats either have a negligible impact in this specific application domain or are inherently mitigated by the system design, with only a few targeted measures required to ensure robustness and reliability.

3.6.2 Transit Residual Detection

Regarding this functionality, each of the aforementioned threats is associated with one (or more) mitigation/control strategies. Each strategy can be marked as *Treated* if already implemented, *Postponed* if it will be addressed later, or *Ignored* if not relevant or applicable to the specific domain under study.

Already treated threats include: *Poisoning* (requiring the use of reliable data sources), *Uniform sampling bias* (addressed through the use of synthetic oversampling techniques, applied only after verifying fairness improvements), and *Data disclosure risk* (ensuring that models comply with differential privacy constraints).

Some mitigation strategies are not relevant for the domain and therefore have been ignored. As an example, a mitigation strategy for data disclosure includes "reducing the information provided by the model," which is not a suitable strategy in this context, given that TrustGuard has been applied to the dataset rather than to the model itself. Similarly, the use of "federated learning to minimize the risk of data breaches" is overly complex for the task at hand, exceeds the available computational resources, and is not relevant to the setting, since the dataset will be released as fully open access, making concerns related to training data breaches inconsistent.

Other threats, although relevant to the task under consideration, have been postponed to align with the development agenda for the dataset and the underlying model improvements. These include conducting a risk analysis for ML applications, performing fairness stress tests under various adversarial or real-world conditions, and integrating ML-specific considerations into the awareness strategy to ensure that all ML stakeholders are adequately informed.

As a future consideration, monitoring logic could be added to identify conditions associated with reduced reliability of residual detection, such as data shifts or atypical input patterns, allowing the system to explicitly signal decreased confidence in the produced results.

3.7 Formal Tool Feedback

TrustSense was effective in raising awareness within the team regarding appropriate practices and procedures for developing trustworthy AI systems. It encouraged reflection on existing workflows and helped identify areas for improvement, contributing to better alignment with trustworthiness principles and fostering more responsible development habits.

The tool was straightforward to use but, at the same time, completing the questionnaire required a certain level of prior understanding of the topics addressed. Some preparation or study was necessary to ensure that responses were well-informed and accurately reflected the team's practices.

TrustGuard proved to be a valuable tool for identifying potential threats related to the AI system and for providing corresponding mitigation recommendations. It supported a structured reflection on risks that may not have been immediately evident during development. The team considers it a useful instrument to be applied in future projects, particularly for improving risk awareness and guiding mitigation planning.

The tool was generally easy to use and facilitated a smooth and guided completion process. However, some questions required advanced technical knowledge about AI-specific security threats and trustworthiness concepts. Moreover, certain frequency-based answer options were too limited, requiring additional clarification through comments.

However, the use of the tool requires additional time, which may not be fully compatible with the deadlines of competitive research groups or industrial contexts. The need to provide a separate TrustGuard analysis for each developed model or dataset can quickly become an overwhelming task, especially given the high degree of similarity among outputs across shared assets and applications.

The Data and AI Model Passports into the Model Hub required some interactions with the developers since there were some technical issues. The tools are considerably useful to document model training metadata within traditional ML/DL pipelines. Certain sections of passport are not currently automatic when considering that the AI models are off-the-shelf. Overall, the tools require a baseline of technical expertise, as the metadata itself assumes a familiarity with core machine learning and deep learning concepts.

4 AI System Evaluation

4.1 Quantitative metrics on pilot data

Several quantitative metrics have been described in D2.1; they are grouped in three main categories: metrics for bias detection in data, metrics for assessing the performance of the AI model, and metrics for evaluating data privacy.

The AI models investigated in the transportation LSP are based both on object detection and classification. Standard detection and classification tasks have been tailored to specific object classes: for example, in Transit Residual Detection, the task is specialized to selected categories of objects (luggage, discarder objects, laptops, clothing, ...), while in Passenger Counting it is focused specifically on people detection. In both cases, detection is performed using bounding boxes rather than segmentation. Based on these settings, the metrics used to assess the performances of the SUT assets comprise the following ones:

- **Box (P R mAP50 mAP50-95)** is a standard grouping of these metrics, and it has been used for both the AI models. Here it follows a description of these metrics:
- **P Precision**, the accuracy of positive predictions. Higher is better: closer to 1 indicates fewer false positives.
- **R Recall**, the ability of the model to detect all relevant instances. Higher is better: closer to 1 means it's capturing more true positives.
- **mAP50** Mean Average Precision at 50% IoU (Intersection over Union, 4th item of table 17 of D2.1) threshold, a common metric for evaluating the accuracy of object detectors. Higher is generally better, indicating more accurate detections: aiming for 0.5 (50%) and above is considered good in many applications, though it heavily depends on the difficulty of the task.
- **mAP50-95** Mean Average Precision averaged over IoU thresholds from 50% to 95% (at 5% steps), providing a more comprehensive view of the model's performance across different levels of detection difficulty. This is a stricter performance metric. Values above 0.3 (30%) can be good, especially on complex datasets.

The Passenger Counting model is tested in real-world scenario, and its performance is assessed using the **mean relative error**, compared with manually annotated ground-truth data, showing an average error of approximately 5.9%, hence achieving an overall accuracy of about 94.1%. Such accuracy has been considered by potential end-users as fully satisfactory, as the model provides reliable quantitative estimates, even if occasional deviations occur. Such deviations have been carefully analysed to fully understand potential and predictable causes, finding that in most cases such deviations are not due to model defects, but to objective conditions (e.g., major occlusions) that make object detection impossible.

The dataset to train the Transit Residual Detection has been designed both to improve the detection and classification performance of the pretrained model (Yolo26n). Along with standard performance metrics used for object detection and classification, a consistency index ("coherence between rotations") has been defined and will be used: this value measures the coherence of predicted labels of an object across its multiple and different views.

At this stage, no other metrics have been considered.

4.2 Trustworthiness Assessment

Prior to the use of the Faith tool *TrustGuard*, a trustworthiness assessment has been performed when drafting and discussing the DPIA for both the passenger counting and the enhanced object detection models. The key point emerging from DPIA is that the LSP on Transportation is the only pilot involving cameras in a public scenario and it is common knowledge that the trust of people being observed with cameras is very difficult to achieve. The presence of video cameras always raises the following issues:

- **Privacy**: AI technology can be used for mass surveillance, allowing authorities or entities to track individuals without their knowledge or consent.
- **Unintended Use**: AI systems deployed for one purpose, such as security or authentication, can be repurposed for other uses without adequate safeguards.
- **Freedom of Expression (Chilling Effect):** People may self-censor or alter their behaviour if they believe they are constantly being monitored, impacting the open and free exchange of ideas.
- **Security**: Systems can store biometric data, and if compromised, this information can be used for identity theft or other malicious purposes.
- **Resiliency**: The system may lack fault tolerance, leaving it unable to predict hardware faults or recover quickly after such events.
- **Biases**: AI algorithms may exhibit biases, leading to inaccurate or discriminatory outcomes, particularly against certain demographic or ethnic groups.
- **Lack of Informed Consent**: Individuals are not informed about or do not provide explicit consent for their faces to be captured and analysed.
- **Misuse by Authorities**: Law enforcement agencies can use AI for surveillance purposes. Without proper regulations, there's a risk of abuse, infringing individuals' rights to privacy and freedom of movement.
- **Accuracy and Reliability**: AI systems may not always be accurate, leading to false positives or negatives. Mistaken identity can have serious consequences, especially in law enforcement applications, where errors can result in unwarranted arrests or other adverse actions.

With these statements in mind, we compiled the following table (Table 17) which shows the priority list of the Trustworthiness attributes in the public transport domain and the possible business consequences.

TW Attributes	Motivation
Privacy-enhanced	Privacy is by far the most important attribute because the trust of people being observed with cameras is very difficult to achieve. *Loss of trust* and *Reputational damage* are the business consequences of a failure.
Secure & Robust	A system which is not secure respect to external attack cause for sure *Loss of trust* & *Reputational damage*
Valid & Reliable	Essential to get accepted from operators and passengers: if not the consequences are Operational inefficiencies and *Loss of trust*.
Accountable & Transparent	The system should be easily controlled for auditing and inspection activities; any choices done should be clearly documented. *Reputational damage* is the danger.
Explainable & Interpretable	These are keys to let the operators using the system to know when it succeeds or when it fails. *Lack of understanding of the system failures* if low on these.
Safe	No dangerous materials are used, and all the electric certifications have been obtained.
Fair	No harmful bias in the chosen set of functionalities.

Table 17 – Prioritization of Trustworthiness attributes in the LSP on public transport.

- High standards of *Privacy* is a must. The consequence of lack of privacy is a complete loss of trust and reputational damage for the company. Details on this parameter in the following paragraph 4.2.1.
- *Security* is straightforward connected with these high standards because a system which is not secure respect to external attacks cause for sure Loss of trust and reputational damage. Insights in paragraph 4.2.2
- *Validation and Reliability* are essential to get accepted from operators and passengers: if not the consequences are inefficiencies and loss of trust. Further details in paragraph 4.2.3.
- Being deployed in a public environment, it is important that the system comply with the *transparency*, thus allowing for auditing and inspection activities. The system should be easily controlled and any choices done should be clearly documented. More information in paragraph 4.2.4.
- *Explainability* is important as well for transparency and system controlling purposes. *Interpretability* is key to let the operators using the system to know when it succeeds or when it fails. More on these two parameters in paragraph 4.2.5.
- *Safety* has a low position in the priority list because we are not using dangerous materials and all the electric certifications have been obtained. No more details are needed.

- *Fairness* is at the bottom of the list because we don't see threats coming from bias in the SUT. There is no additional information on this parameter.

The application of the TrustGuard, supported by the Model and Data Passports, has enabled the evaluation of the AI SUT across key Trustworthy AI characteristics, confirming all the issues and considerations enlightened through the DPIA.

4.2.1 Privacy

Privacy is one of the core vulnerabilities of LSP2 SUT as also pinpointed by TrustGuard results on thread analysis (see Annex I and II). However, when considering the **risk levels**, they are all **low**, since specific mitigation measures were adopted in the functioning and the organization level of the SUT.

In more detail, the Transit Residual Detection service is performed at the end of each service when the bus travels towards the depot or even at the end of each run, but always when there are no passengers; hence, there are no privacy concerns. The processing of passenger data for counting purposes is designed to ensure full protection of individual privacy.

On the other hand, the Passenger Counting functionality installed on board relies on video-based technologies that analyse visual data in real time. However, the processing is carried out using anonymization techniques, so that individuals cannot be identified either directly or indirectly. No personal data is collected, stored, or associated with specific passengers.

The images are processed solely to extract aggregated numerical information (e.g., passenger counts), and any data that could potentially allow identification is either not recorded or immediately transformed into anonymized outputs.

In addition, the integration with vehicle geolocation data does not enable the tracking of individual passengers, as all information remains aggregated and non-attributable to identifiable persons.

Furthermore, within the context of lawful video surveillance, passengers are informed of the presence of such systems through appropriate notices. The use of the service therefore implies awareness and acceptance of the video monitoring conditions; consequently, individuals who do not wish to be recorded may choose not to use the service.

Appropriate technical and organizational safeguards are implemented to ensure that the processing strictly adheres to data minimization and privacy-by-design principles, limiting the use of data exclusively to what is necessary for passenger counting.

4.2.2 Security & Robustness

The system is designed with a comprehensive set of technical and organizational measures to ensure security, integrity, and robustness of processing. Access to the system is restricted through logical access controls, requiring individual user accounts and strong authentication of

credentials. All access and system activities are subject to audit logging to ensure traceability and accountability.

Data transmission is secured through encrypted communication protocols (HTTPS), preventing unauthorized interception. The infrastructure is hosted in a secure data centre environment, with regular backups and redundancy mechanisms in place to ensure business continuity and disaster recovery in case of hardware failures or service disruptions. Backup data is stored in protected and segregated environments, with encryption and access restrictions.

Vulnerability management procedures are implemented, including regular security updates and patching of both infrastructure and application components. Maintenance operations are conducted using controlled procedures, including system snapshots and backup restoration capabilities to preserve data integrity.

The system is designed to minimize risks related to personal data by processing visual inputs only to extract aggregated passenger counts. Images and video streams are not retained beyond the time strictly necessary for processing and are immediately converted into anonymized numerical outputs. This one-way transformation ensures that original data cannot be reconstructed from stored results.

Risk management activities are carried out using established methodologies, and security incidents or data breaches are handled according to formal procedures aligned with applicable data protection regulations. Continuous monitoring and supervision are in place to ensure proper system functioning and to promptly address any technical issues or anomalies.

The AI models and algorithms operate within a controlled environment, without external network communication, reducing exposure to external threats. Updates and improvements are performed in a controlled and manual manner to maintain system reliability and to mitigate potential biases or unforeseen behaviours.

Overall, **the system adopts a privacy-by-design and security-by-design approach, ensuring that data processing remains limited, secure, and resilient against potential risks**.

4.2.3 Validation & Reliability

The Passenger Counting functionality demonstrated a high level of technical maturity and reliability. The validation results confirm that the system remains highly robust across standard transit scenarios. The minor performance fluctuations observed were not due to systemic failures, but rather to real world conditions and/or extreme physical occlusions, such as a passenger carrying balloons directly in front of the camera lens. These instances, while rare, represent the inherent physical limits of computer vision. However, because these errors are easily identifiable via the visual bounding boxes provided to operators, they do not compromise the overall integrity of the data used for aggregate counting. Mitigation strategies, as introduced in the previous section, will be tested during the replication pilot.

For the Transit Residual Detection functionality, the validation phase identified significant concerns regarding the system's reliability in varied operational environments. The results

indicated that the current models did not meet the necessary robustness benchmarks required for high-stakes municipal transit decision-making. To address these deficiencies, the project has launched a series of remedial activities, as detailed in the previous sections. Central to this effort is the creation of a new, sector-specific dataset designed to capture a broader array of edge cases and environmental variables. By retraining the system on this more comprehensive data foundation, the goal is to devise a significantly more robust architecture capable of maintaining consistent performance in a complex, real-world scenario.

4.2.4 Transparency & Accountability

In order to comply with the transparency requirements of the AI Act – specifically the mandate to provide clear and accessible information to users – the Transportation LSP has integrated the use of **Data and AI Model Passports**. These types of standardized documentation serve as the primary mechanism for elucidating the AI lifecycle and ensuring that all stakeholders have a clear view of the system's capabilities, origins, and limitations. By utilizing Data Passports, the project provides a comprehensive audit trail of the information used to train and validate the passenger counting models. These passports document the entire lineage of the data, including its source (the local municipality's transit network), the specific collection methodologies employed, and the preprocessing steps taken. This ensures that data provenance is fully visible, allowing operators and regulators to verify that the information is representative of the urban environment and gathered in accordance with the privacy standards established in the DPIA. Complementing the data documentation, AI Model Passports are used to provide a *technical identity card* for the SUT. These passports detail the model's architecture, its intended use cases, and its performance metrics across various evaluation scenarios (even very challenging, such as high-density crowds or low-light conditions). By documenting the known boundaries and potential failure modes of the SUTs, the Model Passport ensures that transport operators are not dealing with a *black box*. Instead, they are provided with the necessary insights to understand how the AI reaches its conclusions and where human intervention is recommended or necessary.

4.2.5 Explainability & Interpretability

While the system does not utilize dedicated eXplainable AI (XAI) software tools or complex algorithmic heatmaps, the dimension of explainability is addressed through a combination of human-centric training and visual transparency. This approach ensures that the *black box* nature of the AI is mitigated by providing operators with a practical understanding of the system's logic and limitations. Explainability is fostered through a pedagogical approach where developers and transport operators engage in a shared review of system performance. During user training, by performing a joint analysis of errors, the project moves beyond *technical explainability* toward *functional transparency*. Operators are taught not just *that* an error occurred, but *why* it happened—for example, explaining how extreme overcrowding or specific lighting conditions might cause the model to miss a detection. This helps operators build an accurate *mental model* of the AI, allowing them to predict when the system might be less reliable and fostering a deeper

sense of trust. A key technical feature supporting explainability is the *real-time visualization of bounding boxes*. Although primarily a debugging tool, these boxes serve as a primary interface for human oversight. By seeing exactly which objects the AI has identified and framed in the video feed, operators can immediately verify the system's reasoning in real-time. If the system miscounts, the visual evidence (e.g., a bounding box mistakenly framing a large piece of luggage as a person) provides an immediate, intuitive explanation for the error. This transparency allows for rapid identification of false positives or negatives without requiring a background in data science.

4.3 Preliminary Impact Evaluation

The passenger counting asset, for urban and extra-urban buses, can have a significant impact on both operational efficiency and service quality. By providing continuous, fine-grained data on passenger flows, transport operators can better match supply with demand, dynamically adjusting routes, frequencies, and vehicle allocation. This leads to improved load balancing, reduced overcrowding, and more efficient use of fleet resources. In the longer term, the collected data supports strategic planning, enabling evidence-based decisions on network design, timetable optimization, and infrastructure investments.

In addition, when integrated with ticketing or validation systems, passenger counting technologies can support automated consistency checks between the number of detected passengers and validated tickets. This enables the identification of anomalies that may indicate fare evasion, allowing operators to implement more targeted inspection strategies rather than relying solely on random controls. From a user perspective, real-time passenger information can enhance the travel experience by enabling more accurate occupancy predictions and supporting passenger information systems (e.g., indicating crowded vehicles). This contributes to increased comfort, reduced waiting uncertainty, and potentially higher ridership, while also supporting sustainability goals through more efficient operations.

A system for detecting abandoned objects on buses, combined with recognition of litter and dirt, can substantially improve safety, security, and service quality. Automated detection of waste and unclean conditions inside vehicles supports more efficient maintenance and cleaning operations. Instead of relying solely on scheduled or manual inspections, operators can adopt condition-based cleaning strategies, prioritizing vehicles that require immediate attention. This results in better hygiene standards, improved passenger satisfaction, and optimized operational costs. Over time, such systems can also provide insights into usage patterns and behaviour, helping design targeted awareness campaigns and improve overall service quality.

On the other hand, the identification of unattended items may enable rapid intervention by operators or authorities, reducing potential risks and enhancing passenger perception of safety. This is particularly relevant in high-traffic urban environments, where timely responses to suspicious objects are critical.

5 Regulatory and Ethical Assessment

The **Data Protection Impact Assessment (DPIA)** has been conducted in accordance with Article 35 of Regulation (EU) 2016/679, with the aim of identifying, assessing, and mitigating potential risks to the rights and freedoms of individuals arising from the LSP2-SUT. The assessment reflects an accountability-based approach (Article 5(2) GDPR), ensuring that the processing is necessary, proportionate, and compliant with applicable legal frameworks, including Regulation (EU) 2024/1689 and Directive (EU) 2022/2555.

From a regulatory perspective, the processing is based on the legitimate interest of the data controller pursuant to Article 6(1)(f) GDPR. It complies with key principles set out in Article 5 of the GDPR, including data minimization, purpose limitation, and storage limitation. Personal data, initially processed in the form of video streams, is handled in near real time and immediately anonymized, ensuring that outputs consist exclusively of aggregated, non-identifiable data. In this context, the processing aligns with the principles of privacy-by-design and by-default (Article 25 GDPR).

From an ethical standpoint, the system follows an anthropocentric approach, consistent with the principles outlined in the AI Act, ensuring that the deployment of AI-based technologies respects fundamental rights and avoids disproportionate interference with individuals. The processing is strictly limited to LSP2-SUT purposes and does not involve profiling or automated decision-making within the meaning of Article 22 GDPR. Transparency obligations are fulfilled through clear and accessible information notices, in line with Articles 12–13 GDPR.

The DPIA has been developed through a structured process. First, a detailed mapping of data flows was conducted in collaboration with the Data Protection Officer (Articles 37–39 GDPR), identifying how data is collected, processed, and anonymized. Second, the existing security framework of the transport environment was assessed, in line with the requirements of the NIS2 Directive. Third, security measures were integrated with the data processing logic of the system, ensuring consistency with Article 32 GDPR on security of processing. Finally, the DPIA was jointly drafted, consolidating the risk analysis and the identification of appropriate mitigation measures.

The results of the assessment indicate that, following the implementation of the identified technical and organizational measures, the residual risk remains within acceptable thresholds. Therefore, pursuant to Article 36 GDPR, prior consultation with the supervisory authority is not required. Overall, the processing is considered compliant with current regulatory requirements and aligned with ethical principles concerning the protection of individuals' rights and freedoms.

5.1 Alignment with the EU AI Act

The LSP2-SUT has been categorized as a "**low-risk**" system. Despite that, some considerations about the mandates of the AI Act should be made and can be summarized as follows.

Risk Mitigation and Quality Assurance. A fundamental pillar of the AI Act for high-risk applications involves the deployment of an exhaustive risk management framework that spans the entire functional life of the AI tool. This entails a proactive process of spotting, evaluating, and

neutralizing potential hazards that might emerge during initial design, algorithmic training, rollout, or daily operations. This protective cycle must be iteratively refined as the system matures, or as new operational data is gathered. Furthermore, the organizations responsible for these systems must implement rigorous quality assurance protocols to confirm that the AI remains dependable, sturdy, and safe throughout its service life.

In the context of the Transportation LSP, these mandates are met through the application of the FAITH AI Trustworthiness Assessment Framework (AI_TAF). This system offers a methodical approach to governing the risks linked to the LSP2-SUT. By facilitating the orderly detection of threats and the appraisal of vulnerabilities, the framework empowers the local municipal transport authority to pinpoint system weaknesses and apply effective countermeasures. Incorporating these assessment steps into the system's lifecycle ensures a governance model that meets the AI Act's demands for persistent oversight and risk reduction.

Data Governance and Dataset Quality. The AI Act emphasizes the necessity of superior data governance, acknowledging that the integrity and equity of AI outputs are deeply tied to the quality of the underlying training sets. For high-risk systems, datasets used for training and testing must be pertinent, accurately reflect the intended demographic, and remain free from skewed patterns that might impair performance. Moreover, the origin and processing of these datasets must be meticulously recorded to maintain clarity and accountability at every stage of the AI's existence.

Within the Transportation LSP, the local municipality's transit company manages all data assets, ensuring that visual information harvested from the network adheres to high quality benchmarks. The creation and testing of the AI models utilize diverse urban transit data that accounts for various passenger volumes and environmental variables. Crucially, a Data Protection Impact Assessment (DPIA) has been completed to validate that all data processing aligns with privacy laws and protects individual rights. The FAITH project further supports this through Data and Model Passports, which serve as digital ledgers for data lineage, development history, and the privacy safeguards defined in the DPIA, fostering a culture of responsible data stewardship.

Operational Transparency and Human Oversight. To comply with the AI Act, high-risk systems must integrate features that allow for meaningful human supervision, preventing AI results from overriding human expertise in safety-sensitive situations. These oversight mechanisms ensure that trained professionals can interpret what the AI produces, spot inaccuracies, and step in when required. The objective is to position AI as a helpful auxiliary tool rather than a standalone decision-maker in critical public sectors like transport.

The technology evaluated in this Transportation LSP is specifically built as an operational aid, ensuring that the ultimate authority for transit adjustments remains with the human staff. As said, the system has been categorised as "low risk" and has a very limited impact on passengers and operators. The system generates passenger counts and reliability scores that are presented to operators, allowing them to evaluate the model's accuracy within the flow of their work. Transport operators review these findings, if needed, and have the power to validate, adjust, or ignore the AI's data during their management shift. This "human-in-the-loop" strategy

guarantees that while the AI improves efficiency, the municipal authority retains full oversight and accountability.

Documentation Standards. The AI Act also requires that high-risk systems maintain thorough technical records and documentation. These files must provide enough detail for oversight bodies to confirm that the system meets all regulatory benchmarks. Standard documentation covers the technical architecture, training protocols, data characteristics, testing outcomes, and the specific safety measures adopted during the project.

The FAITH initiative provides several tools to help the Transportation LSP maintain these records. The Model Passports offer a standardized summary of the system's traits, its growth stages, and its performance across different transit environments. These are paired with the Data Passports, which reports the data's origin, the municipality's governance methods, and the privacy conditions set by the DPIA. These records, along with internal project reports, create a clear trail of evidence regarding how the system was built and tested, ensuring the level of auditability required for modern high-risk AI applications.

Ethical Considerations in Public Transportation. Beyond just following the law, using AI in public transit brings up vital ethical questions regarding citizen privacy, social equity, and public confidence. AI systems that help optimize transport must be designed with a focus on how they impact data confidentiality and the fair treatment of all commuters. Ethical management in this field goes beyond simple technical metrics; it involves the careful handling of data, openness about how algorithms work, and the active prevention of biases that could unfairly affect specific groups of people.

6 Conclusions and Roadmap for the Final Pilot Phase

6.1 Summary of achievements in the second period

During the initial pilot execution, the Transportation LSP successfully navigated significant operational challenges and achieved key milestones in deploying and assessing its AI-based functionalities. The most notable achievement was the **strategic transition of the SUT from a railway environment to a bus-based public transport network**. This pivot mitigated severe bureaucratic delays associated with Trenitalia and provided a flexible, accessible, and scalable environment for data acquisition and system validation.

The pilot successfully validated the **Passenger Counting functionality, which reached a Technology Readiness Level (TRL) of 6**. Real-world validation demonstrated highly reliable performance with an **average accuracy of 94.1%** (mean relative error of 5.9%). For the **Transit Residual Detection** functionality, initial tests revealed that out-of-the-box models (e.g., YOLO26n) were inadequate for reliably identifying garbage (18%) and unattended luggage(40%). To overcome this, the CNR team successfully **engineered a custom, domain-specific dataset comprising 50,000 images** featuring 300 distinct items. This dataset was constructed using a highly controlled laboratory setup with a rotating platform, multimodal cameras, and advanced generative AI (LanPaint and Juggernaut models) to create photorealistic synthetic environments.

Crucially, the pilot saw the **successful instantiation of the FAITH AI Trustworthiness Assessment Framework (TAF V1)**. The technical teams actively utilized the FAITH toolset:

- **TrustSense** confirmed a moderate-to-high trustworthiness maturity level among the AI developers and operators.
- **Data and AI Model Passports** were successfully implemented on the FAITH AI Model Hub, ensuring end-to-end traceability and transparency regarding data provenance and model architecture.
- **TrustGuard** was used to systematically identify and manage threats. While the system's overall risk profile was confirmed as Low, TrustGuard successfully isolated high-frequency operational threats such as 'Failure/Malfunction' due to environmental variability.

Finally, a comprehensive **Data Protection Impact Assessment (DPIA)** was completed, verifying that the system strictly adheres to privacy-by-design principles. By processing video streams in near real-time at the edge and extracting only anonymized, aggregated numerical data, the SUT successfully aligns with the GDPR and is categorized as a low-risk system under the EU AI Act.

6.2 Plan for Replication Phase

The transition to the bus platform in the initial pilot lays out a highly advantageous foundation for the Replication Phase. Because the alternative bus platform is being already made operative in multiple provinces, **the Replication Pilot will eliminate the need for initial hardware interventions and individual vehicle-by-vehicle installations**. This ensures significantly faster deployment and superior scalability compared to the original railway plan.

The roadmap for the final pilot phase focuses on massive scale-up and algorithmic refinement:

- **Large-Scale Deployment:** The system will be scaled across a massive available fleet consisting of approximately 1,000 vehicles operating across 12 Italian provinces (including Bologna, Modena, Parma, Rimini, and Bari).
- **Model Fine-Tuning:** The custom 50,000-image multi-view dataset will be used to extensively fine-tune the Transit Residual Detection models. The primary goal is to elevate this functionality from its current TRL 3 to the target TRL 5, drastically improving its ability to differentiate between actual garbage, legitimate passenger luggage, and background distractors.
- **Continuous Trustworthiness Monitoring:** The FAITH toolset (TrustGuard, TrustSense, and Passports) will be iteratively applied at this larger scale. Technical countermeasures recommended by TrustGuard, such as drift control mechanisms and input modifications to prevent evasion, will be rigorously evaluated in the expanded operational context.

6.3 Identified risks for the Replication Phase

While the bus environment offers excellent scalability, expanding the SUT to such a large fleet (1,000-vehicle is a very big number) introduces several operational and technical risks that must be managed during the Replication Phase:

- **Data Quality and Environmental Variability:** Public buses suffer from intense environmental perturbations. Variations in lighting, severe passenger crowding, motion blur, and unpredictable physical occlusions (e.g., passengers carrying large items or balloons near the camera lens) pose ongoing risks to consistent detection accuracy.
- **Scalability of System Calibration:** Deploying across 12 provinces involves diverse vehicle layouts (Urban, Suburban, Intercity) and non-homogeneous camera configurations. A significant risk involves the complexity of configuring and calibrating the system across these varying physical environments without extensive manual intervention.
- **Model Drift:** The system is vulnerable to performance degradation over time due to shifts in data distribution, such as changes in passenger clothing (fashion trends) or seasonal variations. Continuous monitoring and periodic retraining will be required to counteract this drift.
- **Human-AI Interaction and Operational Trust:** As the system scales, interacting with a wider array of transport operators increases the risk of human error, misinterpretation of AI outputs, or automation bias (over-reliance on the AI). Ensuring that end-users correctly interpret bounding boxes, understand system limitations, and maintain vigilance remains a critical challenge requiring ongoing, human-centric training.

Mitigation actions concerning the above identified risk, as already described in Section 3.6, are:

- **Data Quality and Environmental Variability** Countermeasures focus on improving robustness under real-world conditions. These include input modification techniques to mitigate sensitivity to noise and environmental variations (e.g., lighting changes, occlusions, camera misalignment), as well as the use of validated and reliable data sources. For the residual detection task, synthetic oversampling is selectively applied to

address sampling bias after verifying that it does not introduce fairness issues. Additionally, future enhancements may include monitoring mechanisms that detect degraded operating conditions and flag reduced confidence in system outputs.

- **Scalability of System Calibration** Calibration is managed through a controlled and relatively static approach, avoiding complex adaptive or continuous retraining strategies. Periodic and planned model updates are preferred over dynamic recalibration, ensuring compatibility with pilot deployment constraints and limited computational resources. More advanced solutions such as federated learning or continuous adaptation have been excluded due to their complexity and lack of proportional benefit in this context.
- **Model Drift** Basic drift control mechanisms are introduced to ensure long-term reliability. These include periodic system-level monitoring and validation checks designed to detect performance degradation over time. While still relatively lightweight, these mechanisms aim to trigger timely interventions when significant drift or data shifts are observed. More advanced drift detection and adaptation strategies are acknowledged but deferred to future development phases.
- **Human-AI Interaction and Operational Trust** Countermeasures emphasize transparency, interpretability, and appropriate human oversight. This includes clearly communicating system limitations, avoiding over-reliance on automated outputs, and ensuring operators understand the conditions under which performance may degrade. Confidence signalling (e.g., flagging low-reliability outputs) is proposed to support informed decision-making. Organisational measures—such as access control, accountability procedures, and alignment with DPIA requirements—further reinforce trust and responsible system use.

7 References[12]

[1] ENISA, 2021. Securing Machine Learning Algorithms. Heraklion: European Union Agency for Cybersecurity

[12] All the references to internet websites are reported in the document as footnotes.

8 Annex I – Passenger Counting Assessment Report

The following pages are the complete assessment report made by the TrustGuard tool related to the Passenger Counting AI Assets:

Model: LSP2-PassengerCounting

Data_ FAITH-LSP2-PassengerCounting

Assessment Report

1. Cartography

Name: Passenger Counting Assessment

Assets: Models: LSP2-PassengerCounting, Data: FAITH-LSP2-PassengerCounting

Lifecycle Stage: PRODUCTION

Criticality Assessment: Medium

Criticality Assessment Questionnaire

Does the AI system falls under the qualification of a high risk AI-system as established in AI Act Annex III or in Article 6 of the AI Act ?

- [x] No
- [ ] Yes

Comment: No comment

Is the AI system used by an Operator of Essential services (OES) and is also used for the provision of an essential service as defined in Annex I of NIS 2 Directive ?

- [x] No
- [ ] Yes

Comment: No comment

Is the AI system used by an Operator of Essential services (OES) and is also used for the provision of an important service as defined in Annex I of NIS 2 Directive ?

- [x] No
- [ ] Yes

Comment: No comment

Is the AI system used for the provision of an important service as defined in Annex I of NIS 2 Directive ?

- [x] No
- [ ] Yes

Comment: No comment

What level of risk does the AI system pose to the health, safety or fundamental rights of natural persons, including influencing the outcome of decision making?

- [ ] Low
- [x] Moderate
- [ ] Substantial
- [ ] High
- [ ] Very High

Comment: No comment

2. Assessment Questionnaire

Threat Assessment

How often does the threat "Evasion" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [x] Once every 10 years
- [ ] Never

> *Comment: Evasion is here considered as a type of attack where an input is intentionally crafted by the attacker in a way that causes the trained model to make a wrong prediction at inference time.*

How often does the threat "Use of adversarial examples crafted in white or grey box conditions" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Oracle" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Poisoning" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Label modification" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Model or data disclosure" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Data disclosure" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Model disclosure" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Compromise of ML application components" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

Comment: No comment

How often does the threat "Failure or malfunction of ML application" occurs in AI systems in this domain?

- [x] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

Comment: No comment

How often does the threat "Human error" occurs in AI systems in this domain?

- [x] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

Comment: No comment

How often does the threat "Denial of service due to inconsistent data or a sponge example" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

Comment: No comment

How often does the threat "Cybersecurity incident not reported to incident response teams" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Prompt attacks" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Propagation of misinformation" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Regulatory and legal non-compliance" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Reduced model interpretability" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Personal video data acquisition" occurs in AI systems in this domain?

- [ ] Once a year
- [x] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

> *Comment: This threat occurs continuously, as the system relies on CCTV cameras to capture video of passengers, resulting in constant acquisition of personal data that is properly managed and protected following GDPR regulations.*

How often does the threat "Training data quality due acquisition procedure" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Bias in Personalized Learning Recommendations" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Model Drift Leading to Inaccurate Assessments" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Misclassification of Structural Defects & Objects" occurs in AI systems in this domain?

- [ ] Once a year
- [x] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

> *Comment: Misclassification of objects occurs with medium frequency in this domain, as challenging conditions such as occlusions, crowd density, and varying lighting can lead the model to incorrectly detect or classify individuals.*

How often does the threat "Dataset Bias from Clear Water Imagery" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Unsafe Route Near Sensitive Structures" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Tampered Mission Parameters" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Opaque Coverage Decisions" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Inspection Neglect Due to Biased Path Sampling" occurs in AI systems in this domain?

- [x] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

> *Comment: This threat occurs with low frequency in this domain, as the system does not perform path sampling, although limited camera coverage and blind spots may lead to partial observation of the environment.*

How often does the threat "No Audit Trail for Route Execution" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Privacy Leakage via Inspection Footage" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

Comment: No comment

How often does the threat "Coverage Failure Due to Invalid Terrain Model" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

Comment: No comment

How often does the threat "Route Logging Exposes Sensitive Infrastructure Layouts" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

Comment: No comment

How often does the threat "Unsafe Classification Leading to Diver Deployment" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

Comment: No comment

How often does the threat "Suboptimal Route Execution Under Uncertain Currents" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Deskilling from lack of transparency" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Diffusion of accountability due to lack of transparency" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Accessible presentation of AI model predictions" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Reduced operator accountability due to lack of interpretability" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Reduced ability of operators to identify prediction errors" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Over-reliance in model predictions" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Security threats - as discussed in TrustGuard assessment" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Environmental risk from AI model error" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Lack of generalizability in production settings" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Lack of generalizability in testing" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Lack of extensive data for training and testing" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Need of external computation resources to infer segmentation" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Lack of a complete documentation on AI model functioning" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Lack of effective explanation methods" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Poor quality of data used in training and testing" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Limited diversity in training data" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

Comment: No comment

How often does the threat "Lack of critical information in training and testing data" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

Comment: No comment

How often does the threat "Lack of robustness checks" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

Comment: No comment

How often does the threat "Unfair discrimination in AI outputs" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

Comment: No comment

How often does the threat "Biased training data" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Inconsistent fairness metrics" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Changes in external conditions" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Discriminatory behavior learned by AI" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Misrepresentation of sensitive characteristics" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Inadequate justification for fairness metrics" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Ineffective data massaging" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Uniform sampling bias" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Reweighing inaccuracies" occurs in AI systems in this domain?

- [x] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

> *Comment: This threat occurs with low frequency in this domain, as reweighing strategies are not commonly a critical component in single-class object detection tasks, although minor inaccuracies may arise in cases of imbalanced training data.*

How often does the threat "Lack of transparency in fairness measures" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Failure to quantify bias" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Unjustified choice of fairness approach" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Dynamic user behavior impacts" occurs in AI systems in this domain?

- [x] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

> *Comment: This threat occurs with medium to high frequency in this domain, as variations in passenger behavior and movement patterns can significantly impact detection and counting performance in real-world conditions.*

How often does the threat "Insufficient safeguards against bias" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Misalignment of fairness objectives" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Inadequate assessment of fairness risks" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Poorly defined fairness thresholds" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Lack of effective fair data representations" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Insufficient documentation of fairness processes" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

Consequence Assessment

What is the business consequence level if the safe operation of the AI system is compromised?

- [ ] Negligible
- [ ] Very Low
- [ ] Low
- [x] Moderate
- [ ] High
- [ ] Very High

> *Comment: The business consequence level is medium, as compromised operation may lead to inaccurate data, operational inefficiencies, and potential regulatory or reputational impacts, although it does not directly affect safety-critical functions.*

What is the business consequence level if the security or the resilience of the AI system to adverse events is compromised?

- [ ] Negligible
- [ ] Very Low
- [ ] Low
- [ ] Moderate
- [x] High
- [ ] Very High

> *Comment: The business consequence level is medium to high, as compromised security or resilience can lead to service disruption, potential data breaches involving sensitive CCTV data, and associated operational, regulatory, and reputational impacts.*

What is the business consequence level if the AI system is not interpretable and explainable?

- [ ] Negligible
- [ ] Very Low
- [x] Low
- [ ] Moderate
- [ ] High
- [ ] Very High

> *Comment: No comment*

What is the business consequence level caused by a privacy compromise of the AI system?

- [ ] Negligible
- [ ] Very Low
- [ ] Low
- [ ] Moderate
- [x] High
- [ ] Very High

> *Comment: No comment*

What is the business consequence level if the AI system is not fair?

- [ ] Negligible
- [ ] Very Low
- [ ] Low
- [x] Moderate
- [ ] High
- [ ] Very High

> *Comment: No comment*

What is the business consequence level if the AI system is not transparent?

- [ ] Negligible
- [ ] Very Low
- [x] Low
- [ ] Moderate
- [ ] High
- [ ] Very High

> *Comment: No comment*

What is the business consequence level if the AI system is not accurate and robust?

- [ ] Negligible
- [ ] Very Low
- [ ] Low
- [x] Moderate
- [ ] High
- [ ] Very High

> *Comment: No comment*

Vulnerability Assessment - LSP2-PassengerCounting

Are input modification techniques (e.g., input sanitisation, randomisation, or noise injection) applied to reduce the risk of attackers exploiting model predictability through controlled input-output pair generation?

- [x] No
- [ ] Yes
- [ ] N/A

> *Comment: No comment*

Have tools or techniques been implemented to detect adversarial inputs or abnormal data during model execution?

- [x] No
- [ ] Yes
- [ ] N/A

> *Comment: No, specific tools or techniques to detect adversarial inputs or abnormal data during model execution have not been implemented, as the system processes controlled video streams and such threats are considered low in this context.*

Is the ML application integrated into the organisation’s security incident detection and response procedures?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Are adversarial training techniques applied during model training to improve robustness against known attack types?

- [x] No
- [ ] Yes
- [ ] N/A

> *Comment: Adversarial attacks are considered low risk in this context.*

Is there an awareness strategy in place that addresses ML-specific attack vectors, and is it actively communicated to all relevant stakeholders involved in model development and deployment?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Has the model architecture been selected and configured with robustness against evasion attacks as a design requirement?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Has the exposure level of the deployed model (e.g., open-source status, API accessibility) been evaluated in relation to attack surface risk?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Have less easily transferable or less commonly exploited model architectures been considered to reduce the risk of adversarial transfer attacks?

- [x] No
- [ ] Yes
- [ ] N/A

> *Comment: No comment*

Is there a security process in place to continuously assess and maintain the security level of ML application components, including model artefacts and infrastructure?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Has the model been configured to limit detailed output (e.g., probability scores, internal logic traces) that could reveal sensitive model characteristics or facilitate adversarial probing?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Have steps been taken to minimise the publicly or externally available information about the model (e.g., architecture, training data, confidence scores) to limit the risk of exploitation?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Do ML applications and their components comply with the organisation's identity management, authentication, and process control policies to ensure secure access and operation?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Has information about the model architecture (e.g., layer types, parameters, training strategies) been restricted to prevent unauthorised access or reverse engineering?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Have differential privacy techniques been implemented during model training or inference to prevent the leakage of sensitive or identifiable information from model outputs?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Are specific poisoning detection or mitigation measures integrated into the workflow after the model evaluation phase, prior to deployment?

- [x] No
- [ ] Yes
- [ ] N/A

Comment: No comment

Are adversarial detection and data sanitisation techniques applied to the training dataset to identify and remove potentially malicious or misleading samples before model training?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Do ML components and their supporting infrastructure adhere to established identity management, authentication, and access control policies to ensure secure and restricted access?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Are there documented and operational processes in place to regularly assess and maintain the security of ML system components (e.g., models, infrastructure, dependencies) throughout their lifecycle?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Are specific indicators of proper model functioning (e.g., accuracy thresholds, drift metrics, anomaly rates) defined and continuously monitored to support early detection of model compromise or malfunction?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Are all external data sources and pretrained models (e.g., for transfer learning) validated for trustworthiness, integrity, and compliance with organisational or regulatory requirements?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Has a comprehensive risk analysis been conducted to identify potential data/model disclosure scenarios across the ML application's lifecycle, including during development, deployment, and operation?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Has federated learning or a comparable privacy-preserving training approach been implemented to reduce the exposure of sensitive data during model training?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Can you ensure that the model cannot reveal information about individual data points?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Is access to detailed information about the model architecture (e.g., layers, parameters, training strategy) restricted to authorised personnel and systems only?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Has the model been configured to limit the detail of its outputs (e.g., suppressing confidence scores, internal states, or debugging information) to reduce the risk of inference or reverse-engineering attacks?

- [x] No
- [ ] Yes
- [ ] N/A

Comment: No comment

Are ML applications integrated into the organisation’s overall security-by-design processes, ensuring that security considerations are addressed throughout all project phases?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Are all third-party and internal components used in the ML pipeline—such as libraries, frameworks, data sources, and infrastructure—regularly assessed for known vulnerabilities and validated for acceptable security levels?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Have existing organisational security policies been updated to explicitly address ML-specific considerations such as model integrity, data provenance, adversarial risks, and lifecycle-specific controls?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Are ML applications regularly reviewed to ensure alignment and compliance with the organisation's established security policies?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Are ML applications formally registered and tracked within the organisation's asset management processes to enable lifecycle oversight and risk management?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Does the model implementation ensure that the model is sufficiently resilient to the environment in which it will operate?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Does the model implementation ensure the inclusion or integration in the AI model a drift control service?

- [x] No
- [ ] Yes
- [ ] N/A

> *Comment: In this context to detect a drift there must be a manual check, so the system cannot automatically monitor if the model becomes less accurate over time.The system relies on the relative stability of the operational environment and periodic manual evaluation.*

Are models designed or selected to support explainability and traceability of outputs, enabling users and auditors to understand the rationale behind decisions?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Has the ML application undergone a compliance assessment to identify and align with applicable laws and regulations (e.g. GDPR, AI Act, sector-specific standards) across all components and lifecycle stages?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Has the model been trained and validated using data and scenarios that accurately reflect the environmental and operational conditions in which it will be deployed?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Are the data and models regularly evaluated using bias detection methods, and are mitigation strategies applied to ensure fairness across relevant demographic or contextual dimensions?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Are ML applications explicitly included in the organisation’s cyber-resilience strategy, with defined procedures for incident response, recovery, and continuity specific to AI systems?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Are AI/ML projects required to meet formal documentation standards covering model design, data sources, training procedures, and intended use?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Are there controls in place to ensure that all data used for training, validation, and inference are sourced, validated, and managed through trusted and verifiable pipelines?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Does the model implementation ensure that Restrict what autonomous agents can do after reading third-party data?

- [ ] No
- [ ] Yes
- [x] N/A

> *Comment: The system does not include autonomous agents or perform actions based on third-party data.*

Does the model implementation ensure that Tag or label data sources to inform the model which parts are safe or unsafe to trust?

- [x] No
- [ ] Yes
- [ ] N/A

> *Comment: No comment*

Is user input clearly separated from system instructions using explicit delimiters or metadata?

- [ ] No
- [ ] Yes
- [x] N/A

> *Comment: The system does not process user-provided inputs or instructions.*

Have the models been fine-tuned to ignore override attempts (e.g., “Ignore previous instructions”)?

- [ ] No
- [ ] Yes
- [x] N/A

> *Comment: The system does not process user-provided inputs or instructions.*

Are mechanisms in place to detect patterns of probing or prompt-revealing questions?

- [ ] No
- [ ] Yes
- [x] N/A

> *Comment: The system does not process user-provided inputs or instructions.*

Are system prompts dynamically changed per session to reduce predictability?

- [ ] No
- [ ] Yes
- [x] N/A

> *Comment: The system does not process user-provided inputs or instructions.*

Is retrieval-augmented generation (RAG) implemented to ground responses in verified and authoritative knowledge sources?

- [ ] No
- [ ] Yes
- [x] N/A

> *Comment: The system does not provide results based on documentation.*

Are entropy-based measures applied to identify uncertainty, instability, or anomalous generation behavior?

- [x] No
- [ ] Yes
- [ ] N/A

> *Comment: No comment*

Is token-level probability estimation used to evaluate the model's confidence in its generated outputs?

- [ ] No
- [ ] Yes
- [x] N/A

> *Comment: The system does not generate textual outputs or tokens.*

Is the origin and lifecycle of the data used to train, fine-tune, or condition the LLM systematically recorded and traceable?

- [ ] No
- [ ] Yes
- [x] N/A

> *Comment: The system does not generate textual outputs or receive textual inputs.*

Are curated and vetted training datasets used to train or fine-tune the models?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Are bias audits and fairness testing systematically conducted on the models and underlying data?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Are bias detection and mitigation techniques implemented, including regular fairness audits of the models and associated artefacts?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Is model performance continuously monitored to detect degradation, drift, or unexpected behavior over time?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Is adaptive path re-planning implemented using real-time current sensors and environmental models?

- [ ] No
- [ ] Yes
- [x] N/A

Comment: No comment

Does the model implementation ensure a human-centred evaluation with operators during production?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Are assessment tools (e.g., uncertainty quantification strategies, outlier detection, and drift detection mechanisms) adopted to detect potential system malfunctions and proactively inform users of possible misbehaviour?

- [x] No
- [ ] Yes
- [ ] N/A

Comment: No comment

Does the model implementation ensure the adoption of privacy preserving strategies and ensure that the data transferred are fully anonimised before sharing on the Net or reduce the computation requirements to ensure the system can work on local premises also in case of low-resource settings?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Are adaptive or online learning techniques employed to handle shifting data distributions?

- [x] No
- [ ] Yes
- [ ] N/A

> *Comment: No comment*

Is regular model re-training scheduled to ensure alignment with current operational realities?

- [x] No
- [ ] Yes
- [ ] N/A

> *Comment: The operational environment and input conditions are relatively stable over time, and updates are performed only when necessary.*

Is model performance continuously monitored in production to detect changes in external conditions?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Are bias regression tests regularly performed during system updates to detect newly introduced disparities?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Are fairness criteria embedded within continuous integration and continuous delivery (CI/CD) pipelines?

- [x] No
- [ ] Yes
- [ ] N/A

> *Comment: Fairness is addressed primarily during data preparation and training.*

Are automated fairness testing tools integrated into the model development lifecycle?

- [x] No
- [ ] Yes
- [ ] N/A

> *Comment: No comment*

Is embedding invariance tested using dedicated fairness evaluation datasets?

- [x] No
- [ ] Yes
- [ ] N/A

Comment: No comment

Are fairness-preserving feature representations generated using learned embeddings?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Are adversarial or contrastive training techniques applied to remove or suppress group identifiers from feature representations?

- [x] No
- [ ] Yes
- [ ] N/A

Comment: No comment

Vulnerability Assessment - FAITH-LSP2-PassengerCounting

Are input modification techniques (e.g., input sanitisation, randomisation, or noise injection) applied to reduce the risk of attackers exploiting model predictability through controlled input-output pair generation?

- [x] No
- [ ] Yes
- [ ] N/A

Comment: No comment

Are appropriate data security controls applied to the dataset throughout its lifecycle, including secure handling, storage, transmission, and access control in the workflows where it is used?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Are adversarial detection and data sanitisation techniques applied to the training dataset to identify and remove potentially malicious or misleading samples before model training?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Are there controls in place to ensure that all data used for training, validation, and inference are sourced, validated, and managed through trusted and verifiable pipelines?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Is the training dataset diversified to improve the model's robustness against data poisoning and ensure it generalises well across varied scenarios?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Are all external data sources and pretrained models (e.g., for transfer learning) validated for trustworthiness, integrity, and compliance with organisational or regulatory requirements?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Has a comprehensive risk analysis been conducted to identify potential data/model disclosure scenarios across the ML application's lifecycle, including during development, deployment, and operation?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Has federated learning or a comparable privacy-preserving training approach been implemented to reduce the exposure of sensitive data during model training?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Is there an awareness strategy in place that addresses ML-specific attack vectors, and is it actively communicated to all relevant stakeholders involved in model development and deployment?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Have differential privacy techniques been implemented during model training or inference to prevent the leakage of sensitive or identifiable information from model outputs?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Can you ensure that the model cannot reveal information about individual data points?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Does the model implementation ensure that the model is sufficiently resilient to the environment in which it will operate?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Has the model been trained and validated using data and scenarios that accurately reflect the environmental and operational conditions in which it will be deployed?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Has the ML application undergone a compliance assessment to identify and align with applicable laws and regulations (e.g. GDPR, AI Act, sector-specific standards) across all components and lifecycle stages?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Are the data and models regularly evaluated using bias detection methods, and are mitigation strategies applied to ensure fairness across relevant demographic or contextual dimensions?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Are bias audits and fairness testing systematically conducted on the models and underlying data?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Is data provenance systematically tracked to ensure traceability?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Are legal counsel or privacy experts engaged to review data sources and their associated terms of service?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Are imaging data processed without being stored, with only alerts (i.e., generated outputs) and relevant metadata transmitted to the control station?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Are encryption mechanisms, role-based access controls, and redaction of sensitive visual elements applied before data storage or sharing?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Is there a data policy or mechanism to ensure the adoption of data-quality assessment tools to evaluate the data used during the training and testing phases?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

3. Assessment Results

Trustworthiness Property: Accountable and transparent

Threat: Model or data disclosure

Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Compromise of ML application components
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Compromise of ML application components
Risk Score: Low
Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Model or data disclosure
Risk Score: Low
Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Human error
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Trustworthiness Property: Fair with harmful bias managed

Threat: Failure or malfunction of ML application
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Regulatory and legal non-compliance
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Bias in Personalized Learning Recommendations
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Insufficient safeguards against bias
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions:

Mitigation Control: Embed fairness criteria into continuous integration/continuous delivery pipelines.
Status: Ignored

> *Comment: In this context, the risk is considered low*

Mitigation Control: Use automated fairness testing tools within the model development lifecycle.
Status: Ignored

> *Comment: In this context, the risk is considered low*

Threat: Changes in external conditions
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions:

Mitigation Control: Employ adaptive or online learning techniques to handle shifting data distributions.
Status: Ignored

> *Comment: The use of adaptive or online learning techniques is not considered necessary in this context, as the operational environment and input conditions are relatively stable over time. Model updates are instead performed in a controlled manner when required.*

Mitigation Control: Schedule regular model re-training to align with current realities.

Status: Ignored

> *Comment: The use of adaptive or online learning techniques is not considered necessary in this context, as the operational environment and input conditions are relatively stable over time. Model updates are instead performed in a controlled manner when required.*

Threat: Lack of effective fair data representations
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions:

Mitigation Control: Test embedding invariance using fairness evaluation datasets.

Status: Ignored

> *Comment: The use of automated fairness testing tools is not considered necessary in this context, as the system performs aggregate object detection and counting rather than individual-level decision-making.*

Mitigation Control: Apply adversarial or contrastive training to remove group identifiers from features.

Status: Ignored

> *Comment: The use of automated fairness testing tools is not considered necessary in this context, as the system performs aggregate object detection and counting rather than individual-level decision-making.*

Threat: Poisoning
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Regulatory and legal non-compliance
Risk Score: Low
Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Failure or malfunction of ML application
Risk Score: Low
Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Poisoning
Risk Score: Low
Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Trustworthiness Property: Safe

Threat: Failure or malfunction of ML application
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Compromise of ML application components
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Evasion
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions:

Mitigation Control: Use adversarial training techniques to train the AI model

Status: Ignored

> *Comment: In this context, the risk is considered low.*

Mitigation Control: Implement tools to detect if a data point is an adversarial example or not

Status: Ignored

> *Comment: In this context, the risk is considered low.*

Threat: Poisoning
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions:

Mitigation Control: Integrate poisoning control after the "model evaluation" phase

Status: Treated

> *Comment: Poisoning control is already addressed after model evaluation through additional validation and data integrity checks prior to deployment, ensuring that the model has not been adversely influenced by corrupted or unreliable training data*

Threat: Failure or malfunction of ML application
Risk Score: Low
Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Poisoning
Risk Score: Low
Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Trustworthiness Property: Secure and resilient

Threat: Denial of service due to inconsistent data or a sponge example
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Data disclosure
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Model or data disclosure
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Use of adversarial examples crafted in white or grey box conditions
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Compromise of ML application components
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Oracle
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Evasion
Risk Score: Low
Asset: LSP2-PassengerCounting

Mitigation Actions:

Mitigation Control: Use less easily transferable models
Status: Ignored

> *Comment: In this context, the risk is considered low*

Mitigation Control: Use adversarial training techniques to train the AI model
Status: Ignored

> *Comment: In this context, the risk is considered low.*

Mitigation Control: Implement tools to detect if a data point is an adversarial example or not
Status: Ignored

> *Comment: In this context, the risk is considered low.*

Mitigation Control: Apply modifications to inputs
Status: Ignored

> *Comment: In this context, the risk is considered low*

Threat: Poisoning
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions:

Mitigation Control: Integrate poisoning control after the "model evaluation" phase
Status: Treated

> *Comment: Poisoning control is already addressed after model evaluation through additional validation and data integrity checks prior to deployment, ensuring that the model has not been adversely influenced by corrupted or unreliable training data*

Threat: Compromise of ML application components
Risk Score: Low
Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Evasion
Risk Score: High
Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions:

Mitigation Control: Apply modifications to inputs
Status: Postponed

> *Comment: No comment*

Threat: Denial of service due to inconsistent data or a sponge example
Risk Score: Low
Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Data disclosure
Risk Score: Low
Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Model or data disclosure
Risk Score: Low
Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Poisoning
Risk Score: Low
Asset: FAITH-LSP2-PassengerCounting

Mitigation Actions: No mitigation actions are required

Threat: Model disclosure
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions:

Mitigation Control: Reduce the output information given by the model

Status: Ignored

> *Comment: Model disclosure is inherently addressed by limiting the system outputs to aggregated results without exposing detailed model information or internal parameters.*

Trustworthiness Property: Explainable and interpretable

Threat: Propagation of misinformation
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions:

Mitigation Control: Apply retrieval-augmented generation (RAG).

Status: Ignored

> *Comment: This threat does not apply in this system as it doesn't accept input text or provide output text.*

Mitigation Control: Introduce entropy-based measures.

Status: Ignored

> *Comment: This threat does not apply in this system as it doesn't accept input text or provide output text.*

Mitigation Control: Estimate token-level probability.

Status: Ignored

> *Comment: This threat does not apply in this system as it doesn't accept input text or provide output text.*

Mitigation Control: Recording and trace the origin and lifecycle of the data used to train, fine-tune, or condition the LLM.

Status: Ignored

> *Comment: This threat does not apply in this system as it doesn't accept input text or provide output text.*

Threat: Failure or malfunction of ML application
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Reduced ability of operators to identify prediction errors
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Insufficient safeguards against bias
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions:

Mitigation Control: Embed fairness criteria into continuous integration/continuous delivery pipelines.

Status: Ignored

> *Comment: In this context, the risk is considered low*

Mitigation Control: Use automated fairness testing tools within the model development lifecycle.

Status: Ignored

> *Comment: In this context, the risk is considered low*

Threat: Reduced operator accountability due to lack of interpretability
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Failure or malfunction of ML application
Risk Score: Low
Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Human error
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Trustworthiness Property: Privacy enhanced

Threat: Data disclosure
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Model or data disclosure
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Use of adversarial examples crafted in white or grey box conditions
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Oracle
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Need of external computation resources to infer segmentation
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Poisoning
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions:

Mitigation Control: Integrate poisoning control after the "model evaluation" phase

Status: Treated

> *Comment: Poisoning control is already addressed after model evaluation through additional validation and data integrity checks prior to deployment, ensuring that the model has not been adversely influenced by corrupted or unreliable training data*

Threat: Personal video data acquisition
Risk Score: Low
Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Privacy Leakage via Inspection Footage
Risk Score: Low
Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Data disclosure
Risk Score: Low

Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Model or data disclosure
Risk Score: Low
Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Poisoning
Risk Score: Low
Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Model disclosure
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions:

Mitigation Control: Reduce the output information given by the model

Status: Ignored

> *Comment: Model disclosure is inherently addressed by limiting the system outputs to aggregated results without exposing detailed model information or internal parameters.*

Trustworthiness Property: Valid and reliable

Threat: Denial of service due to inconsistent data or a sponge example
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Propagation of misinformation
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions:

Mitigation Control: Estimate token-level probability.

Status: Ignored

> *Comment: This threat does not apply in this system as it doesn't accept input text or provide output text.*

Mitigation Control: Apply retrieval-augmented generation (RAG).

Status: Ignored

> *Comment: This threat does not apply in this system as it doesn't accept input text or provide output text.*

Mitigation Control: Introduce entropy-based measures.

Status: Ignored

> *Comment: This threat does not apply in this system as it doesn't accept input text or provide output text.*

Mitigation Control: Recording and trace the origin and lifecycle of the data used to train, fine-tune, or condition the LLM.

Status: Ignored

> *Comment: This threat does not apply in this system as it doesn't accept input text or provide output text.*

Threat: Lack of generalizability in production settings
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions:

Mitigation Control: Adopt assessment tools to detect possible malfunctioning of the system (e.g., uncertainty quantification strategies, outliers detection, drift detection tools) and inform the user when there can be potential misbehaviour of the system

Status: Postponed

> *Comment: It is planned to adopt assessment tools to detect possible malfunctioning of the system*

Threat: Failure or malfunction of ML application
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions:

Mitigation Control: Include or integrate in the AI model a drift control service

Status: Postponed

> *Comment: It is planned to include or integrate in the AI model a drift control service*

Threat: Use of adversarial examples crafted in white or grey box conditions
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Compromise of ML application components
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Prompt attacks
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions:

Mitigation Control: Fine-tune models to ignore override attempts (e.g., "Ignore previous instructions").

Status: Ignored

> *Comment: The system does not provide text outputs or text inputs.*

Mitigation Control: Separate user input from system instructions using clear delimiters or metadata.

Status: Ignored

> *Comment: The system does not provide text outputs or text inputs.*

Mitigation Control: Restrict what autonomous agents can do after reading third-party data

Status: Ignored

> *Comment: The system does not provide text outputs or text inputs.*

Mitigation Control: Tag or label data sources to inform the model which parts are safe or unsafe to trust.

Status: Ignored

> *Comment: The system does not provide text outputs or text inputs.*

Mitigation Control: Change system prompts per session to reduce predictability.

Status: Ignored

> *Comment: The system does not provide text outputs or text inputs.*

Mitigation Control: Detect patterns of probing or prompt-revealing questions.

Status: No action taken

> *Comment: No comment*

Threat: Model Drift Leading to Inaccurate Assessments
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Changes in external conditions
Risk Score: Low

Asset: LSP2-PassengerCounting
Mitigation Actions:

Mitigation Control: Employ adaptive or online learning techniques to handle shifting data distributions.

Status: Ignored

> *Comment: The use of adaptive or online learning techniques is not considered necessary in this context, as the operational environment and input conditions are relatively stable over time. Model updates are instead performed in a controlled manner when required.*

Mitigation Control: Schedule regular model re-training to align with current realities.

Status: Ignored

> *Comment: The use of adaptive or online learning techniques is not considered necessary in this context, as the operational environment and input conditions are relatively stable over time. Model updates are instead performed in a controlled manner when required.*

Threat: Lack of effective fair data representations
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions:

Mitigation Control: Test embedding invariance using fairness evaluation datasets.

Status: Ignored

> *Comment: The use of automated fairness testing tools is not considered necessary in this context, as the system performs aggregate object detection and counting rather than individual-level decision-making.*

Mitigation Control: Apply adversarial or contrastive training to remove group identifiers from features.

Status: Ignored

> *Comment: The use of automated fairness testing tools is not considered necessary in this context, as the system performs aggregate object detection and counting rather than individual-level decision-making.*

Threat: Poisoning
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions:

Mitigation Control: Integrate poisoning control after the "model evaluation" phase

Status: Treated

> *Comment: Poisoning control is already addressed after model evaluation through additional validation and data integrity checks prior to deployment, ensuring that the model has not been adversely influenced by corrupted or unreliable training data*

Threat: Compromise of ML application components
Risk Score: Low
Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Poor quality of data used in training and testing
Risk Score: Low
Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Denial of service due to inconsistent data or a sponge example
Risk Score: Low
Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Failure or malfunction of ML application
Risk Score: Low
Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Poisoning
Risk Score: Low

Asset: FAITH-LSP2-PassengerCounting
Mitigation Actions: No mitigation actions are required

Threat: Suboptimal Route Execution Under Uncertain Currents
Risk Score: Low
Asset: LSP2-PassengerCounting
Mitigation Actions:

Mitigation Control: Implement adaptive path re-planning using real-time current sensors and environment models

Status: Ignored

> *Comment: This measure is not applicable in this context, as the system does not perform navigation or route execution, but operates on fixed camera inputs for passenger detection and counting.*

9 Annex II -Transit Residual Detection Assessment Report

The following pages are the complete assessment report made by the TrustGuard tool related to the Transit Residual Detection AI Asset:

Data: LSP2-TransitResidualDetection

Assessment Report

1. Cartography

Name: Transit Residual Detection Assessment

Assets: Data: LSP2-TransitResidualDetection

Lifecycle Stage: TESTING

Criticality Assessment: Low

Criticality Assessment Questionnaire

Does the AI system falls under the qualification of a high risk AI-system as established in AI Act Annex III or in Article 6 of the AI Act ?

- [x] No
- [ ] Yes

Comment: No comment

Is the AI system used by an Operator of Essential services (OES) and is also used for the provision of an essential service as defined in Annex I of NIS 2 Directive ?

- [x] No
- [ ] Yes

Comment: No comment

Is the AI system used by an Operator of Essential services (OES) and is also used for the provision of an important service as defined in Annex I of NIS 2 Directive ?

- [x] No
- [ ] Yes

Comment: No comment

Is the AI system used for the provision of an important service as defined in Annex I of NIS 2 Directive ?

- [x] No
- [ ] Yes

Comment: No comment

What level of risk does the AI system pose to the health, safety or fundamental rights of natural persons, including influencing the outcome of decision making?

- [x] Low
- [ ] Moderate
- [ ] Substantial
- [ ] High
- [ ] Very High

Comment: No comment

2. Assessment Questionnaire

Threat Assessment

How often does the threat "Evasion" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

Comment: No comment

How often does the threat "Use of adversarial examples crafted in white or grey box conditions" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [x] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

Comment: No comment

How often does the threat "Oracle" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

Comment: No comment

How often does the threat "Poisoning" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [x] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

Comment: No comment

How often does the threat "Label modification" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [x] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

> *Comment: No comment*

How often does the threat "Model or data disclosure" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Data disclosure" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Model disclosure" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Compromise of ML application components" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

Comment: No comment

How often does the threat "Failure or malfunction of ML application" occurs in AI systems in this domain?

- [ ] Once a year
- [x] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

Comment: No comment

How often does the threat "Human error" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

Comment: No comment

How often does the threat "Denial of service due to inconsistent data or a sponge example" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

Comment: No comment

How often does the threat "Cybersecurity incident not reported to incident response teams" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Prompt attacks" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Propagation of misinformation" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Regulatory and legal non-compliance" occurs in AI systems in this domain?

- [ ] Once a year
- [x] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

> *Comment: No comment*

How often does the threat "Reduced model interpretability" occurs in AI systems in this domain?

- [ ] Once a year
- [x] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

> *Comment: No comment*

How often does the threat "Personal video data acquisition" occurs in AI systems in this domain?

- [x] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

> *Comment: No comment*

How often does the threat "Training data quality due acquisition procedure" occurs in AI systems in this domain?

- [ ] Once a year
- [x] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

> *Comment: No comment*

How often does the threat "Bias in Personalized Learning Recommendations" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Model Drift Leading to Inaccurate Assessments" occurs in AI systems in this domain?

- [ ] Once a year
- [x] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

> *Comment: No comment*

How often does the threat "Misclassification of Structural Defects & Objects" occurs in AI systems in this domain?

- [ ] Once a year
- [x] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

> *Comment: No comment*

How often does the threat "Dataset Bias from Clear Water Imagery" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Unsafe Route Near Sensitive Structures" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Tampered Mission Parameters" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Opaque Coverage Decisions" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Inspection Neglect Due to Biased Path Sampling" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "No Audit Trail for Route Execution" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Privacy Leakage via Inspection Footage" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Coverage Failure Due to Invalid Terrain Model" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Route Logging Exposes Sensitive Infrastructure Layouts" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Unsafe Classification Leading to Diver Deployment" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Suboptimal Route Execution Under Uncertain Currents" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

Comment: No comment

How often does the threat "Deskilling from lack of transparency" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

Comment: No comment

How often does the threat "Diffusion of accountability due to lack of transparency" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

Comment: No comment

How often does the threat "Accessible presentation of AI model predictions" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

Comment: No comment

How often does the threat "Reduced operator accountability due to lack of interpretability" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Reduced ability of operators to identify prediction errors" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Over-reliance in model predictions" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [x] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

> *Comment: No comment*

How often does the threat "Security threats - as discussed in TrustGuard assessment" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Environmental risk from AI model error" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Lack of generalizability in production settings" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Lack of generalizability in testing" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Lack of extensive data for training and testing" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Need of external computation resources to infer segmentation" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Lack of a complete documentation on AI model functioning" occurs in AI systems in this domain?

- [ ] Once a year
- [x] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

> *Comment: No comment*

How often does the threat "Lack of effective explanation methods" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Poor quality of data used in training and testing" occurs in AI systems in this domain?

- [x] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

> *Comment: No comment*

How often does the threat "Limited diversity in training data" occurs in AI systems in this domain?

- [x] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

Comment: No comment

How often does the threat "Lack of critical information in training and testing data" occurs in AI systems in this domain?

- [x] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

Comment: No comment

How often does the threat "Lack of robustness checks" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [x] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

Comment: No comment

How often does the threat "Unfair discrimination in AI outputs" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [x] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

Comment: No comment

How often does the threat "Biased training data" occurs in AI systems in this domain?

- [x] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

> *Comment: No comment*

How often does the threat "Inconsistent fairness metrics" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Changes in external conditions" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [x] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

> *Comment: No comment*

How often does the threat "Discriminatory behavior learned by AI" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Misrepresentation of sensitive characteristics" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Inadequate justification for fairness metrics" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Ineffective data massaging" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Uniform sampling bias" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Reweighing inaccuracies" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Lack of transparency in fairness measures" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Failure to quantify bias" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Unjustified choice of fairness approach" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Dynamic user behavior impacts" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Insufficient safeguards against bias" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [x] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [ ] Never

> *Comment: No comment*

How often does the threat "Misalignment of fairness objectives" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Inadequate assessment of fairness risks" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Poorly defined fairness thresholds" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Lack of effective fair data representations" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

How often does the threat "Insufficient documentation of fairness processes" occurs in AI systems in this domain?

- [ ] Once a year
- [ ] Twice a year
- [ ] Once every 2 years
- [ ] Once every 5 years
- [ ] Once every 10 years
- [x] Never

> *Comment: No comment*

Consequence Assessment

What is the business consequence level if the safe operation of the AI system is compromised?

- [ ] Negligible
- [x] Very Low
- [ ] Low
- [ ] Moderate
- [ ] High
- [ ] Very High

> *Comment: No comment*

What is the business consequence level if the security or the resilience of the AI system to adverse events is compromised?

- [ ] Negligible
- [x] Very Low
- [ ] Low
- [ ] Moderate
- [ ] High
- [ ] Very High

> *Comment: No comment*

What is the business consequence level if the AI system is not interpretable and explainable?

- [ ] Negligible
- [ ] Very Low
- [ ] Low
- [x] Moderate
- [ ] High
- [ ] Very High

> *Comment: No comment*

What is the business consequence level caused by a privacy compromise of the AI system?

- [ ] Negligible
- [ ] Very Low
- [ ] Low
- [ ] Moderate
- [x] High
- [ ] Very High

> *Comment: No comment*

What is the business consequence level if the AI system is not fair?

- [x] Negligible
- [ ] Very Low
- [ ] Low
- [ ] Moderate
- [ ] High
- [ ] Very High

> *Comment: No comment*

What is the business consequence level if the AI system is not transparent?

- [ ] Negligible
- [ ] Very Low
- [x] Low
- [ ] Moderate
- [ ] High
- [ ] Very High

> *Comment: No comment*

What is the business consequence level if the AI system is not accurate and robust?

- [ ] Negligible
- [ ] Very Low
- [x] Low
- [ ] Moderate
- [ ] High
- [ ] Very High

> *Comment: No comment*

Vulnerability Assessment - LSP2-TransitResidualDetection

Are all external data sources and pretrained models (e.g., for transfer learning) validated for trustworthiness, integrity, and compliance with organisational or regulatory requirements?

- [x] No
- [ ] Yes
- [ ] N/A

> *Comment: No comment*

Are there controls in place to ensure that all data used for training, validation, and inference are sourced, validated, and managed through trusted and verifiable pipelines?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Are adversarial detection and data sanitisation techniques applied to the training dataset to identify and remove potentially malicious or misleading samples before model training?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Are appropriate data security controls applied to the dataset throughout its lifecycle, including secure handling, storage, transmission, and access control in the workflows where it is used?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Is the training dataset diversified to improve the model's robustness against data poisoning and ensure it generalises well across varied scenarios?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Has a comprehensive risk analysis been conducted to identify potential data/model disclosure scenarios across the ML application's lifecycle, including during development, deployment, and operation?

- [x] No
- [ ] Yes
- [ ] N/A

Comment: No comment

Is there an awareness strategy in place that addresses ML-specific attack vectors, and is it actively communicated to all relevant stakeholders involved in model development and deployment?

- [x] No
- [ ] Yes
- [ ] N/A

Comment: No comment

Has federated learning or a comparable privacy-preserving training approach been implemented to reduce the exposure of sensitive data during model training?

- [x] No
- [ ] Yes
- [ ] N/A

Comment: No comment

Have differential privacy techniques been implemented during model training or inference to prevent the leakage of sensitive or identifiable information from model outputs?

- [ ] No
- [ ] Yes
- [x] N/A

Comment: No comment

Can you ensure that the model cannot reveal information about individual data points?

- [x] No
- [ ] Yes
- [ ] N/A

Comment: No comment

Are imaging data processed without being stored, with only alerts (i.e., generated outputs) and relevant metadata transmitted to the control station?

- [ ] No
- [ ] Yes
- [x] N/A

Comment: No comment

Is a subsequent testing phase (Phase II) planned in which the train operates in motion without passengers (e.g., along a decoy route) to validate system performance under realistic but controlled conditions?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Has a synthetic dataset (e.g., generative or mixed-reality) been developed to incorporate more realistic data for model training and testing?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Is a dedicated DEBUG mode defined that, under appropriate and controlled conditions, allows temporary local video storage for troubleshooting and validation purposes?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Are environmental models regularly updated?

- [x] No
- [ ] Yes
- [ ] N/A

Comment: No comment

Is there a data policy or mechanism to ensure the adoption of privacy preservation measures that enable keeping track of relevant information about sensible attributes in the clinical data?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Are data balancing techniques (e.g., under-sampling or over-sampling) applied to mitigate potential bias in the data?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Are the data sourced from diverse origins to ensure representation of all relevant population subgroups?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Are the data collection and preprocessing methods transparently documented?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Are sample distributions analyzed for hidden skews or imbalances prior to model training?

- [ ] No
- [x] Yes
- [ ] N/A

Comment: No comment

Is synthetic oversampling applied only after verifying that it leads to measurable fairness improvements?

- [x] No
- [ ] Yes
- [ ] N/A

Comment: No comment

Is stratified or proportional sampling applied to ensure adequate representation of relevant subgroups?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Are bias trends visualized over time or across model versions to monitor changes?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Are group fairness metrics (e.g., demographic parity, equalized odds) computed on a regular basis?

- [ ] No
- [ ] Yes
- [x] N/A

> *Comment: No comment*

Are fairness metrics evaluated side-by-side with accuracy metrics to assess potential trade-offs?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Are system-level risk assessments conducted that explicitly include fairness-related impacts?

- [ ] No
- [x] Yes
- [ ] N/A

> *Comment: No comment*

Is the risk register regularly updated to document identified fairness-related threats?

- [x] No
- [ ] Yes
- [ ] N/A

> *Comment: No comment*

Are fairness stress tests performed under adversarial or realistic operational conditions?

- [x] No
- [ ] Yes
- [ ] N/A

> *Comment: No comment*

3. Assessment Results

Trustworthiness Property: Accountable and transparent

Threat: Model or data disclosure
Risk Score: Low
Asset: LSP2-TransitResidualDetection
Mitigation Actions:

Mitigation Control: Conduct a risk analysis of the ML application

Status: Postponed

> *Comment: No comment*

Threat: Compromise of ML application components
Risk Score: Low
Asset: LSP2-TransitResidualDetection
Mitigation Actions:

Mitigation Control: Conduct a risk analysis of the ML application

Status: Postponed

> *Comment: No comment*

Trustworthiness Property: Fair with harmful bias managed

Threat: Poisoning
Risk Score: Low
Asset: LSP2-TransitResidualDetection
Mitigation Actions:

Mitigation Control: Ensure reliable sources are used

Status: Treated

> *Comment: No comment*

Threat: Lack of critical information in training and testing data
Risk Score: Low
Asset: LSP2-TransitResidualDetection
Mitigation Actions: No mitigation actions are required

Threat: Biased training data
Risk Score: Low
Asset: LSP2-TransitResidualDetection
Mitigation Actions: No mitigation actions are required

Threat: Uniform sampling bias
Risk Score: Low
Asset: LSP2-TransitResidualDetection
Mitigation Actions:

Mitigation Control: Use synthetic oversampling only after testing for fairness improvements.

Status: Treated

> *Comment: No comment*

Threat: Failure to quantify bias
Risk Score: Low
Asset: LSP2-TransitResidualDetection
Mitigation Actions:

Mitigation Control: Regularly compute group fairness metrics (e.g., demographic parity, equalized odds).

Status: Treated

> *Comment: No comment*

Threat: Inadequate assessment of fairness risks
Risk Score: Low
Asset: LSP2-TransitResidualDetection
Mitigation Actions:

Mitigation Control: Regularly update the risk register with identified fairness threats.

Status: Ignored

> *Comment: No comment*

Mitigation Control: Perform fairness stress tests under various adversarial or real-world conditions.

Status: Postponed

> *Comment: No comment*

Trustworthiness Property: Safe

Threat: Poisoning
Risk Score: Low
Asset: LSP2-TransitResidualDetection
Mitigation Actions:

Mitigation Control: Ensure reliable sources are used

Status: Treated

> *Comment: No comment*

Trustworthiness Property: Secure and resilient

Threat: Poisoning
Risk Score: Low
Asset: LSP2-TransitResidualDetection
Mitigation Actions:

Mitigation Control: Ensure reliable sources are used

Status: Treated

> *Comment: No comment*

Threat: Model or data disclosure
Risk Score: Low
Asset: LSP2-TransitResidualDetection
Mitigation Actions:

Mitigation Control: Conduct a risk analysis of the ML application

Status: Postponed

> *Comment: No comment*

Threat: Data disclosure
Risk Score: Low
Asset: LSP2-TransitResidualDetection
Mitigation Actions:

Mitigation Control: Integrate ML specificities to awareness strategy and ensure all ML stakeholders are receiving it

Status: Postponed

> *Comment: No comment*

Mitigation Control: Ensure that models respect differential privacy

Status: Treated

> *Comment: No comment*

Mitigation Control: Reduce the information given by the model

Status: Ignored

> *Comment: No comment*

Threat: Compromise of ML application components
Risk Score: Low
Asset: LSP2-TransitResidualDetection
Mitigation Actions:

Mitigation Control: Conduct a risk analysis of the ML application

Status: Postponed

> *Comment: No comment*

Trustworthiness Property: Explainable and interpretable

Threat: Lack of critical information in training and testing data
Risk Score: Low
Asset: LSP2-TransitResidualDetection
Mitigation Actions: No mitigation actions are required

Trustworthiness Property: Privacy enhanced

Threat: Poisoning
Risk Score: Low
Asset: LSP2-TransitResidualDetection
Mitigation Actions: No mitigation actions are required

Threat: Model or data disclosure
Risk Score: Low
Asset: LSP2-TransitResidualDetection
Mitigation Actions:

Mitigation Control: Conduct a risk analysis of the ML application

Status: Postponed

> *Comment: No comment*

Threat: Data disclosure
Risk Score: Low
Asset: LSP2-TransitResidualDetection
Mitigation Actions:

Mitigation Control: Use federated learning to minimise the risk of data breaches

Status: Ignored

> *Comment: No comment*

Mitigation Control: Integrate ML specificities to awareness strategy and ensure all ML stakeholders are receiving it

Status: Postponed

> *Comment: No comment*

Mitigation Control: Ensure that models respect differential privacy

Status: Treated

> *Comment: No comment*

Mitigation Control: Reduce the information given by the model

Status: Ignored

> *Comment: No comment*

Threat: Personal video data acquisition
Risk Score: Very High
Asset: LSP2-TransitResidualDetection
Mitigation Actions:

Mitigation Control: Image data acquired are not stored, only processed, and only alerts (following outputs) along with metadata are sent to the control station.

Status: Postponed

> *Comment: No comment*

Threat: Lack of critical information in training and testing data
Risk Score: Low
Asset: LSP2-TransitResidualDetection
Mitigation Actions: No mitigation actions are required

Trustworthiness Property: Valid and reliable

Threat: Poisoning
Risk Score: Low
Asset: LSP2-TransitResidualDetection
Mitigation Actions:

Mitigation Control: Ensure reliable sources are used

Status: Treated

> *Comment: No comment*

Threat: Compromise of ML application components
Risk Score: Low
Asset: LSP2-TransitResidualDetection
Mitigation Actions:

Mitigation Control: Conduct a risk analysis of the ML application

Status: Postponed

> *Comment: No comment*

Threat: Training data quality due acquisition procedure
Risk Score: Low
Asset: LSP2-TransitResidualDetection
Mitigation Actions: No mitigation actions are required

Threat: Coverage Failure Due to Invalid Terrain Model
Risk Score: Low
Asset: LSP2-TransitResidualDetection
Mitigation Actions:

Mitigation Control: Update environmental models regularly

Status: Ignored

> *Comment: No comment*