Ref. Ares(2026)5892247 - 10/06/2026

# D10.2

Report initial pilot - AI in active ageing solutions

| | |
|---|---|
| **Related Work Package** | **WP10 – Implementation of Active Ageing LSP** |
| **Related Task** | Task 10.2. – Initial Pilot |
| **Lead Beneficiary** | AOA |
| **Contributing Beneficiaries** | BRIDG |
| **Document version** | V.1.0 |
| **Deliverable Type** | Report |
| **Distribution Level** | PU |
| **Contractual Date of Delivery** | 30/04/2026 |
| **Actual Date of Delivery** | 10/06/2026 |



| | |
|---|---|
| **Authors** | **Pilar Sala (AOA), Juan Carlos Naranjo (AOA)** |
| **Contributors** | Gloria Cea (BRD), Alba Gallego (BRD), Alfredo Sanz (BRD) |
| **Reviewers** | Alberto Topini (UNIFI), Panagiota Papageorgiou (KUL) |



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

## Version History

| Version | Description | Date completed |
|---|---|---|
| 0.1 | ToC | 21/01/2026 |
| 0.2 | First draft | 15/03/2026 |
| 0.3 | Intermediate version | 22/04/2026 |
| 0.4 | Version for internal review | 27/05/2026 |
| 0.5 | Refinements considering the internal review suggestions | 07/06/2026 |
| 1.0 | Ready for submission | 09/06/2026 |

## Statement of Originality

This deliverable contains original, unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and 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 the full publication of this document.

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

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

## Executive summary

This deliverable reports the progress, preliminary results, and main lessons learned from the first pilot of the AI-based active ageing solution implemented within LSP7. The pilot has moved from preparation to real-world deployment in Valencia, where a privacy-preserving home monitoring system is being used to support ageing in place by unobtrusively observing daily activity patterns and generating meaningful information for relatives and caregivers. In parallel with the technical deployment, the pilot has also addressed user engagement, trust-building, and trustworthiness assessment, in line with the broader objectives of the FAITH project.

At the time of reporting, the solution had been deployed in 55 homes, with 175 active users in total: 55 older adults and 120 relatives or informal caregivers. Recruitment and onboarding were supported through training sessions, co-creation workshops, digital outreach, and collaboration with local stakeholders and social services. The results of these activities show generally positive perceptions of AI in elderly care, while also confirming that privacy remains a central concern and that clear communication, explanation, and continuous support are essential for adoption and sustained use.

The pilot has also enabled the structured application of the FAITH AI_TAF v1 methodology and associated tools, including trustSense, trustGuard, and the AI Model Hub. Using these instruments, the LSP partners have conducted cartography, threat analysis, impact assessment, vulnerability assessment, risk analysis, and countermeasure definition for the AI system under test. This has provided a formal, traceable trustworthiness profile for the solution, covering dimensions such as privacy, reliability, interpretability, fairness, safety, transparency, and accountability, and demonstrating the practical value of embedding trust assessment in a real deployment context.

The first pilot has also revealed several operational lessons highly relevant to the replication phase. The most significant deviation from the original plan was the delay in obtaining ethical approval, which affected the start of data collection and slowed early recruitment. To address this, the consortium expanded outreach through digital campaigns and community networks, thereby increasing uptake over time. Other lessons include the importance of early and continuous stakeholder engagement, phased onboarding, privacy-by-design communication, and progressing toward closer-to-real-time data processing to improve the usefulness of system outputs for caregivers.

The report D10.2 confirms that a privacy-preserving, AI-enabled home monitoring solution can be successfully implemented in a real active aging environment. It also demonstrates that a trustworthiness assessment can be effectively integrated into the pilot execution in a structured and meaningful manner. The evidence gathered during this phase establishes a strong foundation for the replication stage, during which the solution will be expanded to additional sites and user groups. Meanwhile, the evaluation framework and governance approach developed in the first pilot will continue to guide the deployment and ongoing improvements.

## Table of Contents

1 Introduction ..... 8
1.1 Scope and objectives of the 1st pilot in the Active Ageing LSP ..... 8
1.2 Structure of the document ..... 9
2 Progress on the AI system deployment(s) and execution of the Pilots ..... 10
2.1 Status of the 1st Pilot execution ..... 10
2.1.1 Deployment Context & Domain Specifics ..... 10
2.1.2 Training ..... 11
2.1.3 Users ..... 11
2.1.4 Evaluation Process ..... 12
2.2 Status of the Replication Pilot preparation/initialization ..... 12
2.3 Deviations from the original deployment plan ..... 13
2.4 Data collection throughout the 1st Pilot execution ..... 14
2.5 Integration into LSP workflows: Lessons learned ..... 17
3 Application of FAITH Tools ..... 19
3.1 Application of the FAITH AI_TAF V1 methodology ..... 19
3.1.1 Cartography ..... 19
3.1.2 Threat Analysis ..... 21
3.1.3 Impact Assessment ..... 22
3.1.4 Vulnerability Assessment ..... 23
3.1.5 Risk Analysis ..... 25
3.1.6 Countermeasures ..... 26
3.2 Tool Validation Planning & Training Processes ..... 28
3.2.1 Dry-run ..... 28
3.2.2 Plan development ..... 29
3.2.3 Tool Training ..... 32
3.3 FAITH Tool Validation Activities ..... 33
3.3.1 trustSense ..... 33
3.3.2 Data & AI model Passport in the FAITH AI Model Hub ..... 35
3.3.3 trustGuard ..... 37
3.4 Threat Analysis ..... 39
3.5 Initial Risk Assessment results ..... 41

3.6 Countermeasures ..... 42
4 AI System Evaluation ..... 44
4.1 Quantitative metrics on pilot data ..... 44
4.1.1 Sociodemographic characterisation of participants ..... 44
4.1.2 AI System performance ..... 45
4.1.3 Intermediate measurement of AI trust and perceptions ..... 50
4.2 Trustworthiness Assessment ..... 52
4.2.1 Validation & Reliability ..... 52
4.2.2 Security & Robustness ..... 53
4.2.3 Transparency & Accountability ..... 53
4.2.4 Explainability & Interpretability ..... 54
4.2.5 Safety ..... 54
4.2.6 Fairness ..... 55
4.2.7 Privacy ..... 55
4.3 Preliminary Impact Evaluation ..... 56
5 Regulatory and Ethical Assessment ..... 57
5.1 Data Protection Pathway and Certification ..... 57
5.2 Alignment with the EU AI Act and Ethical Considerations ..... 60
6 Conclusions and roadmap for the final pilot phase ..... 63
6.1 Summary of achievements in the second period ..... 63
6.2 Plan for Replication Phase ..... 64
6.3 Identified risks for the Replication Phase ..... 65
7 References ..... 67

## List of Abbreviations

| Abbreviation | Explanation |
|---|---|
| **ADL** | Activity of Daily Living |
| **AES** | Advanced Encryption Standard |
| **AI** | Artificial Intelligence |
| **AI_TAF** | Artificial Intelligence Trustworthiness Assessment Framework |
| **AHA** | Active and Healthy Ageing |
| **AOA** | Activage.org |
| **API** | Application Programming Interface |
| **AWS** | Amazon Web Services |
| **CFS** | Clinical Frailty Scale |
| **DBCV** | Density-Based Clustering Validation. |
| **DoA** | Description of Action |
| **DPIA** | Data Protection Impact Assessment |
| **DPO** | Data Protection Officer |
| **DVC** | Data Version Control |
| **EIA** | Ethical Impact Assessments |
| **EU** | European Union |
| **GDP** | Gross Domestic Product |
| **GDPR** | General Data Protection Regulation |
| **HCT** | Human-Computer Trust |
| **HDBSCAN** | Hierarchical Density-Based Spatial Clustering of Applications with Noise |
| **ICT** | Information and Communication Technology |
| **IFU** | Instructions For Use |
| **KPI** | Key Performance Indicators |
| **LSP** | Large Scale Pilot |
| **ML** | Machine Learning |
| **RBAC** | Role-Based Access Control |
| **RDS** | Relational Database Service |
| **REST** | Representational State Transfer |
| **SHAP** | SHapley Additive exPlanations |
| **STM** | System Trust Modeller |
| **SUT** | System Under Test |
| **TLS** | Transport Layer Security |
| **ToC** | Table of Contents |
| **TRL** | Technology Readiness Level |
| **TiA** | Trust-in-Automation |
| **TXAI** | XAI Trust Scale |
| **UC** | Use Case |
| **UI/UX** | User Interface/User Experience |
| **UMAP** | Uniform Manifold Approximation and Projection |
| **UNESCO** | United Nations Educational, Scientific and Cultural Organization |
| **WP** | Work Package |

## List of Tables

Table 1. Plan development · planning activities ..... 29
Table 2. Plan development · training and support activities ..... 30
Table 3. Plan development · FAITH tools deployment and iterative implementation ..... 31
Table 4. Plan development · Trust assessment and monitoring activities ..... 31
Table 5. Plan development · AI Model Hub and platform integration activities ..... 32
Table 6. Cluster identification ..... 45
Table 7. Risks for replication phase ..... 65

## List of Figures

Figure 1. Session 1 of workshop “Vivir mejor con IA confinable” ..... 11
Figure 2 Dagster ETL pipeline implementation steps ..... 15
Figure 3 Phases of FAITH AI_TAF V1 methodology ..... 19
Figure 4 trustSense AI technical team result ..... 34
Figure 5. LSP7 Data Passport ..... 36
Figure 6. LPS7 AI Model passport ..... 37
Figure 7. Fist page LSP7 trustGuard Assesment ..... 38
Figure 8 Distribution users in clusters ..... 46
Figure 9 Transition between clusters ..... 47
Figure 10 Cluster classification per date ..... 48
Figure 11 Fitted Function Slow-Slow OWL ..... 48
Figure 12 Fitted Function Slow-Fast Earlybird ..... 49
Figure 13 Fitted Function Fast-Fast Hummingbird ..... 49
Figure 14 Fitted Function Fast-Slow Sargeant ..... 50
Figure 15 Concern levels ..... 51
Figure 16 AI system perception ..... 52

# 1 Introduction

## 1.1 Scope and objectives of the 1st pilot in the Active Ageing LSP

This deliverable addresses the first pilot phase of LSP7, in which the AI-based solution is introduced in a real-life active ageing setting, end users and stakeholders are engaged, trustworthiness assessment activities are launched, and first evidence is gathered on usability, operational viability, and consistency with the FAITH framework. The findings from this stage are intended to support the subsequent replication phase and the further consolidation of the active ageing use case.

The scope of the first pilot in the Active Ageing LSP covers the deployment and early validation of an AI-enabled home monitoring solution designed to support older adults ageing in place in real domestic environments. More specifically, the pilot focuses on the use of privacy-preserving sensing technologies and AI functionalities for the observation and interpretation of daily activity patterns, with the purpose of generating meaningful information for relatives and caregivers while respecting the privacy expectations associated with home-based care.

In addition to assessing the technical operation of the solution, the first pilot also considers the wider trustworthiness dimensions of the AI System Under Test, in line with the goals of the FAITH project. This includes the application of the FAITH AI_TAF methodology and associated tools to examine system risks, organisational preparedness, governance practices, transparency, and trust-related aspects linked to the deployment of AI in the active ageing domain.

The main objectives of the first pilot are the following:

- Deployment of the AI System Under Test in a real-life home and care support environment within the Active Ageing LSP
- Verification of the operational feasibility of the solution under realistic use conditions involving older adults, relatives, and caregivers.
- Assessment of user engagement, acceptability, and practical suitability through training, onboarding, and interaction with the deployed system.
- Evaluation of key trustworthiness dimensions, including privacy, reliability, interpretability, fairness, safety, and accountability.
- Application of FAITH tools such as trustSense, trustGuard, and the AI Model Hub to support structured trustworthiness assessment and governance activities
- Collection of operational evidence, user perceptions, and lessons learned to support refinement of the solution and preparation of the replication phase.
- Preparation for the extension of the pilot to additional deployment settings and user groups in the next phase of LSP7.

Through these objectives, the first pilot establishes the initial evidence base for the trustworthy use of AI-supported home monitoring solutions in active ageing settings.

## 1.2 Structure of the document

This report is organised to follow the progression of the first pilot from deployment activities to assessment outcomes and next-step planning. After this introductory section, the document first examines how the pilot has been implemented in practice and then moves towards the trustworthiness analysis, evaluation findings, and regulatory reflections derived from that implementation.

**Chapter 2** concentrates on the actual execution of the pilot in the Active Ageing setting. It describes the operational context in Valencia, the involvement of older adults and caregivers, the training and engagement actions carried out, the way the pilot has been evaluated so far, the status of the replication preparation, and the main adjustments made during implementation, together with the initial evidence collected and the lessons emerging from integration into the LSP workflow.

**Chapter 3** shifts the focus to the use of FAITH methods and supporting tools. It explains how the FAITH AI_TAF v1 methodology has been applied in LSP7 and how instruments such as trustSense, trustGuard, and the AI Model Hub have contributed to the identification of assets, the analysis of threats and vulnerabilities, the estimation of risks, and the definition of initial mitigation actions.

**Chapter 4** presents the pilot evaluation results with attention to both technical and trustworthiness-related evidence. In this part, the document reports the available quantitative indicators, reviews the assessment of the main trustworthiness dimensions, and outlines the first observations on the effects and relevance of the solution in the active ageing domain.

**Chapter 5** addresses the compliance-oriented perspective of the pilot. It discusses the main ethical and regulatory issues associated with the system under test, including its possible path towards recognised conformity routes and its positioning with respect to the EU AI Act and other applicable obligations.

Finally, **Chapter 6** brings together the main outcomes of the reporting period and connects them with the next phase of work. It summarises the progress achieved during the initial pilot and sets out the basis for the replication stage, including the forward plan and the risks that still need to be managed.

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

## 2.1 Status of the 1st Pilot execution

The 1st pilot of the Active Ageing LSP has progressed from the preparatory stage to operational execution in Valencia, where the AI system under test is currently being deployed and assessed under real-life conditions involving older adults and their caregivers. At this stage, the main organisational, ethical, and technical conditions for pilot execution have been put in place, enabling the effective rollout of the solution in the target environment.

The pilot concerns the deployment of Vitana as a privacy-preserving home monitoring solution intended to support ageing in place through the unobtrusive monitoring of daily activity patterns. In parallel, recruitment and onboarding activities have continued through existing user networks, digital outreach actions, and collaboration with local stakeholders and social services in Valencia. At the time of reporting, 55 homes had already been equipped with the solution and 120 relatives had been involved, corresponding to a total of 175 users in operation.

### 2.1.1 Deployment Context & Domain Specifics

The 1st pilot is situated in the active ageing domain and, more specifically, in the context of home-based monitoring to support older adults living alone or spending most of the day alone in their homes. Within this setting, the AI system under test is deployed as a privacy-preserving monitoring solution intended to support ageing in place through the continuous observation and analysis of daily activity patterns, while providing relevant information to relatives and caregivers.

The deployment takes place in a sensitive socio-technical environment characterised by the involvement of older adults, informal caregivers, service providers, and public-sector stakeholders, as well as by the need to ensure ethical compliance, data protection, and user trust. In contrast to intrusive monitoring approaches, the deployed solution relies on passive sensing technologies, including infrared, door, and key sensors, and explicitly avoids the use of audio or video capture, thereby aligning the technical design of the pilot with the privacy expectations associated with elderly care in domestic environments.

From a domain perspective, the relevance of the pilot is linked to the growing need for solutions that can support independent living, early risk detection, and more timely interventions in the context of demographic ageing and increasing pressure on care systems. The pilot therefore addresses a use case in which behavioural segmentation and anomaly detection may contribute to identifying changes in daily routines, mobility, sleep patterns, or room occupancy that could be associated with frailty progression or emerging health and wellbeing risks. In this respect, the deployment context is not limited to technical validation, but also concerns the practical suitability, acceptability, and trustworthiness of AI-enabled monitoring in real-life home environments.

### 2.1.2 Training

Training and awareness-raising activities were carried out to support end-user familiarisation with the pilot and to prepare participants for the operational use of the system. In particular, an end-user training conference was organised at the Senior University of Valencia, with 150 attendees, as part of the broader recruitment, onboarding, and user engagement process.

In parallel, two co-creation workshops were held on 14 January 2026 as part of the pilot outreach activities. These workshops served not only as instruments for recruitment, but also as mechanisms for trust-building and expectation management, helping participants to better understand the purpose, scope, and implications of the system under test in the context of active ageing.



*Figure 1. Session 1 of workshop "Vivir mejor con IA confinable"*

The training and workshop activities also supported the collection of an initial baseline assessment regarding trustworthiness perceptions. Out of the 150 participants involved in the training session, 45 completed the Wooclap questionnaire, corresponding to a 30% response rate. The available results indicate generally positive expectations towards the use of AI in elderly care, while also highlighting privacy as a key concern and confirming the need for clear communication, explanation, and user-oriented support throughout pilot execution.

### 2.1.3 Users

The initial pilot is addressed primarily to older adults aged 65 and over, especially those living alone or spending most of the day alone at home, and to their relatives and informal caregivers, who act as the main secondary users of the system. In operational terms, the deployed solution is already being used by 175 users in total, comprising 55 older adults and 120 relatives, which confirms that both primary and secondary user groups are actively involved in the pilot execution.

From a user-profile perspective, the available evidence indicates that the pilot population includes older adults and family members with heterogeneous but generally adequate levels of

digital readiness for participation in the study. The workshop and training data show that most respondents were between 66 and 75 years old, with 52 identifying as users and 32 as family members, while approximately 82% reported medium-to-high technological familiarity. At the same time, 64% indicated no prior experience with care technologies, which confirms the relevance of training, clear communication, and continuous support as enabling conditions for adoption and effective use.

The user context is therefore characterised by a dual perspective: older adults as the main beneficiaries and monitored users of the solution, and relatives or caregivers as the main interpreters of alerts, activity information, and AI-supported insights. This distribution is fully aligned with the intended use of the system, which combines ageing-in-place support for older adults with remote supervision and decision support for caregivers in everyday care situations.

### 2.1.4 Evaluation Process

The evaluation process of the initial pilot has been carried out through a combination of deployment, user engagement, and evidence collection activities performed under real-life conditions. First, the sensor-based home monitoring infrastructure and the cloud processing environment were deployed and validated in operational conditions, enabling continuous installation, onboarding, live data collection, and system use in the target environment. In parallel, behavioural segmentation was integrated as a core AI functionality, while anomaly detection advanced through iterative validation and refinement activities.

User-centred evaluation actions were also implemented during the pilot execution. Two co-creation workshops were held on 14 January 2026 as part of the outreach and engagement process, and workshops with end users were conducted during January to collect a baseline assessment of trustworthiness perceptions regarding the system under test. In addition, an end-user training conference was organised at the Senior University of Valencia with 150 attendees, and 45 participants completed the Wooclap questionnaire, providing an initial structured dataset on expectations, privacy concerns, and perceived trust conditions related to the use of AI in elderly care.

The evaluation process also included the collection of operational and technical evidence from the deployed solution. Real-time data processing from live users was enabled for model ingestion, cluster predictions were connected to the application, and caregivers were able to visualise model outputs in daily operation. At the same time, the Vitana Web Portal was operative, the caregiver app was available in public app stores, and the AI-enhanced application was released to an internal testing group, allowing the pilot to gather evidence on software operability and AI integration in practice.

## 2.2 Status of the Replication Pilot preparation/initialization

The replication pilot of the Active Ageing LSP is planned to extend the reach of LSP7 beyond the initial deployment site in Valencia, gradually expanding to other municipalities and regions in Spain, with potential for broader EU-scale implementation in subsequent phases. In accordance with the Description of Action, the replication phase runs from M20 to M51, targeting the involvement of an additional 250 users — elderly individuals and their caregivers — and

integrating an advanced AI model for anomaly detection as well as the FAITH AI Trustworthiness Assessment Framework (FAITH AI_TAF) and the System Trust Modeller (STM).

At the current stage, preparatory activities for the replication pilot are underway. Discussions have been held with the Valencia Regional Government and municipal social services to define the geographic scope and operational conditions for the second deployment phase. The municipality of Alaquàs has been identified as a candidate site for initial replication activities, as it offers a relevant expansion context, access to elderly populations, and public administration support. These stakeholder-level engagements constitute the first step in establishing the organisational and institutional framework required for a successful replication rollout.

Stakeholder consultation activities have provided initial evidence of interest in geographic expansion. A session was held in Madrid with 8 participants from a professional women's retiree association to assess interest and identify potential barriers to engaging user communities beyond Valencia. Furthermore, three sessions were held with service providers (8 participants) and four sessions with representatives from public administration (10 participants) to explore replication involvement and align expectations regarding the deployment model, data governance arrangements, and service delivery responsibilities. These consultations have also helped identify at least five service providers whose collaboration is expected to support the recruitment process in replication areas.

From a technical standpoint, preparatory work has focused on defining the architecture for using foundational models in the AI SUT's anomaly detection component to address the potential challenge of data scarcity at new sites during the early stages of the replication deployment. This planning activity addresses the recognised limitation that current anomaly detection models require a sufficient volume of user-level behavioural data to generate reliable, personalised outputs, a condition that may not be fully met at the outset of deployment in new locations. The adoption of foundational model approaches is expected to mitigate this risk and support more rapid model adaptation to new user cohorts.

Regarding user onboarding and recruitment, the replication strategy follows the same multi-channel framework employed in the initial pilot, involving digital outreach, senior centre engagement, and collaboration with local social service networks. The foreseen participant profile remains consistent with the initial pilot: older adults aged 65 and over living alone or spending most of the day at home, and their relatives or informal caregivers. The replication pilot will also incorporate the refined trustworthiness assessment methodology developed and validated during the initial pilot phase, ensuring continuity in the evaluation approach across both deployment phases.

## 2.3 Deviations from the original deployment plan

The Active Ageing LSP has, in general, followed the deployment rationale defined in the FAITH Description of Action, namely the progressive deployment and validation of an AI-powered, privacy-preserving home monitoring solution for elderly individuals in Valencia, with a subsequent replication phase extending to additional sites. The core elements of the intended approach, sensor-based monitoring of activities of daily living, AI-based behavioural

segmentation and anomaly detection, involvement of older adults and their caregivers, and evaluation of trustworthiness using the FAITH AI_TAF, have all been maintained and operationalised as planned.

The main deviation from the original deployment schedule concerned the timeline for obtaining formal ethical approval. The process of obtaining clearance from the Ethics Committee of the Universitat Politècnica de València (UPV) took longer than initially foreseen, causing a delay at the start of the pilot data collection activities. Ethical approval was eventually granted, enabling the formal commencement of the initial pilot, but the delay had a downstream impact on the pace of user recruitment and the start of operational data collection. As a result, the volume of active users onboarded during the pilot's early months fell short of initial projections.

To address the slower-than-expected recruitment pace, the consortium implemented a set of corrective actions. Paid digital campaigns were launched through social media channels to broaden awareness and attract new participants beyond the existing Vitana user base. Outreach activities were also expanded through engagement with a wider range of elderly associations and community organisations in Valencia, complementing the existing channels through the Valencia Sandbox ecosystem and social services collaboration. These measures proved effective in progressively increasing user enrolment, as reflected by the figure of 55 elderly users and 120 caregivers — a total of 175 users — actively using the deployed solution at the time of reporting.

A further adaptation concerned the operational model for the AI-enhanced application. The Vitana AI-Enhanced App, which integrates AI-based anomaly detection and behavioural segmentation into a caregiver-facing interface, was still in internal testing at the time of reporting and had not been released to the full pilot user base. This reflects the consortium's iterative development approach, which prioritises validation and quality assurance before broader release. The core monitoring infrastructure, the caregiver portal, and the mobile application are fully operational and in use by pilot participants, while the AI-enhanced application continues to mature through live internal testing.

These deviations did not affect the core purpose or evaluation objectives of the initial pilot. The data collection infrastructure has been validated in operational conditions, the behavioural segmentation model has been integrated as a functional AI service, and the FAITH AI_TAF methodology has been applied through trustGuard, trustSense, and the AI Model Hub. The ethical approval delay and the adjusted recruitment trajectory are common operational challenges in pilots involving vulnerable populations and sensitive personal data, and the corrective measures implemented have ensured that the pilot is now progressing in line with its intended objectives. These experiences also provide valuable practical guidance for planning and managing the replication phase.

## 2.4 Data collection throughout the 1st Pilot execution

Data collection in the first pilot iteration has been organised along two parallel streams: continuous IoT sensor data generated by the home monitoring infrastructure deployed across participating homes, and structured questionnaire data collected from participants at defined measurement points. All collection activities are governed by the ethical approval granted by the

Ethics Committee of the Universitat Politècnica de València (CEI/UPV), and comply with GDPR and the Spanish Organic Law 3/2018 [1](LOPDGDD), with explicit informed consent as the legal basis for processing personal data. Given that some participants may be in situations of dependency, consent was assessed in this pilot as being freely given, with participation remaining voluntary and without any adverse consequences for care or support. Participants are assigned a pseudonymous alphanumeric identifier at enrolment, with the mapping to real identity held exclusively in a locked paper file accessible only to the principal investigator.

The IoT continuously collects data during LSP7's 1st Pilot execution, including daily activities of senior citizens at home. The activity is collected through various sensors: motion sensors in key locations (Bedroom, Living Room, Kitchen, and Bathroom), a door sensor installed at the home's entrance, and a tag sensor attached to the keychain that detects when the senior citizen has left their home.

The data is collected in real time as timestamped sensor events, such as motion detection or door opened/closed. However, this raw data is not ready for AI consumption as is; it must be heavily pre-processed. The pre-processing pipeline consists of two main steps: extracting a time series from the point-like events and then fitting a parametric family of functions to each day's activity data to obtain a feature vector of fitted parameters. These are the feature vectors consumed by the clustering model and provide a solid basis for representing behavioural patterns of elderly people.



*Figure 2 Dagster ETL pipeline implementation steps*

The ETL pipeline is implemented in Dagster[2], an asset-centric data processing orchestrator that supports data partitioning (e.g., by user or date) and easy tracking of successfully processed data and potential failure states. It is divided into 5 substeps as shown in the figure. First (raw_event_data), the event data collected by the sensors for a given day is queried via the VITANA API. Then (claned_event_data), only the necessary data is extracted from the events: the sensor ID, which can be directly mapped to a location in the home; a unique ID for each home or elderly user; and the event timestamp. After cleaning all these events, a final filter selects the events that will contribute (selected_events_for_timeseries), and the time series is constructed by assuming the elderly person stays at the location where it was last detected (daily_timeseries). The result is a regularly sampled time series (as opposed to the arbitrarily spaced timestamps from the events) of home activity that starts at 00:00 of a given day (roughly when the user goes to bed) and goes until 08:00 on the next day (roughly when the user gets out of bed). That is, a 32-hour window centred around the main day activity with a sampling resolution of 1 minute. Finally, a function is fitted to the activity in order to characterise it. Currently, all the locations

[1] https://www.boe.es/eli/es/lo/2018/12/05/3/con
[2] https://dagster.io/

are being worked on, but only the bedroom is fully fledged. For the bedroom (bedroom_funcfit), the fitting function is the sum of two step functions in the shape of a "well" (see section 4.1), with a total of 7 trainable parameters. Once the fitted parameters are obtained, they are stored in a PostgreSQL database for later access, such as when the AI-enhanced app shares the clustering model's insights with caregivers.

Data is currently processed in batches, and efforts are underway to move it as close to real-time as possible, especially for the Replication Pilot. Ideally, as soon as all the data for the day is gathered, the processing would be automatically kick-started

At the time of reporting, the system had been deployed in 55 homes (55 elderly users and 120 relatives/caregivers, for a total of 175 active users). Data collection started from the first installations in January 2025, accumulating approximately 9,000 sensor events, with the collection window extending to February 2026 as deployments continued for this report.

Structured questionnaire data were collected from participants both elderly and relatives, including sociodemographic profiles, technology literacy levels, and clinical frailty scores for elderly participants.

In addition, XAI Trust Scale was conducted as part of the co-creation and awareness sessions held on 14 January 2026 at the Senior University of Polytechnic University Valencia, using the Wooclap interactive platform. These sessions were designed primarily for outreach, trust-building and onboarding, but also served as an opportunity to collect an early post-exposure assessment of trustworthiness perceptions.

The XAI Trust Scale[3] is a psychometrically grounded instrument specifically designed to measure user trust in AI systems that produce or rely on explanations. Items probe core trust dimensions directly relevant to XAI contexts: confidence in the system, perceived predictability, reliability, efficiency, and believability of outputs, dimensions that capture both trust-in-a-prediction (moment-level trust of individual outputs) and trust-in-a-model (sustained belief that the AI performs reliably on real-world data). The scale is explicitly designed as a repeat measure, reflecting the processual nature of trusting in AI systems and the expectation that trust states evolve — including toward contingent, justified, stable trust or, under conditions of high false-alarm rates or conspicuous errors, toward rapid breakdown. This property is particularly relevant to the FAITH LSP7 pilot context, where family caregivers and formal care workers interact with an IoT-based daily activity monitoring system that generates AI-driven anomaly alerts for elderly people living alone. In this setting, trust is not formed once but recalibrated continuously as users encounter alerts, assess their accuracy, and decide whether to act on them; research confirms that, in AI health monitoring for older adults, the credibility and actionability of system outputs are primary determinants of whether caregivers adopt or abandon a technology.

[3] Hoffman, R. M., Mueller, S. T., & Klein, G. (2021). Measuring trust in the XAI context. PsyArXiv. https://doi.org/10.31234/osf.io/e3kv9

## 2.5 Integration into LSP workflows: Lessons learned

The initial pilot of the Active Ageing LSP has provided the first practical evidence on how the Vitana-based AI monitoring solution integrates into the real-life workflows of older adults and informal caregivers. While the primary objective of this phase has been the deployment and initial validation of the AI system in isolation, the process of embedding the solution into real-life workflows has also surfaced a number of challenges and insights that are directly relevant to the continuation of the pilot and the preparation for the replication phase.

The most structurally significant issue encountered during the initial pilot was the delay in obtaining formal ethical clearance from the Ethics Committee of the Universitat Politècnica de València (CEI/UPV). This delay postponed the start of formal data collection and the full activation of user onboarding workflows. This experience underlines the importance of initiating ethical approval procedures well in advance of intended deployment, particularly when the pilot involves sensitive personal data from a vulnerable population.

The physical installation of the sensor infrastructure across participating homes introduced logistical challenges specific to the residential setting. Variations in home layouts, available connectivity, and older adults' prior digital experience required the technical team to adapt installation procedures on a case-by-case basis. In several instances, additional visits were needed to ensure that sensors were functioning correctly and that the participants felt comfortable with the installed equipment. These early findings confirm the operational complexity of deploying IoT-based monitoring in real domestic environments and highlight the need for flexible, well-documented field procedures that can accommodate the heterogeneity of residential settings within the pilot population.

An important integration challenge emerged regarding how informal caregivers interact with system outputs in their everyday routines. Although the AI-Enhanced Vitana app is still in internal testing, the caregiver-facing mobile application and the Vitana Web Portal were operational and in use by the 120 relatives involved in the pilot. Preliminary feedback indicated that some users needed additional support to correctly interpret the monitoring insights. In particular, the transition from conventional forms of checking in with elderly relatives towards technology-mediated activity monitoring required a period of familiarisation and trust-building.

The experience of the initial pilot has generated a number of lessons that are directly relevant to how the Active Ageing LSP is organised and executed going forward. The following points are considered the most significant:

- **Early and continuous stakeholder engagement is essential for sustainable integration.** The co-creation workshops conducted in January 2026 demonstrated that direct, face-to-face engagement is the most effective mechanism for building trust, managing expectations, and preparing older adults and their caregivers for the use of AI-enabled monitoring in their everyday lives. These activities should be maintained and expanded in both the remaining initial pilot period and the replication phase.
- **Phased and supported onboarding reduces early dropout and sustains participation.** The combination of recruitment outreach, in-person installation support, and follow-up

engagement through trusted intermediaries proved important for sustaining participant involvement beyond the initial enrolment stage. The evidence available at the time of reporting, with 55 elderly users and 120 caregivers actively using the solution, confirms that when adequate support is provided, older adults and their relatives are willing and able to engage with the pilot over time.

- **Privacy-preserving design is a prerequisite for user acceptance in the home monitoring context.** The explicit avoidance of audio and video capture, the use of passive sensing technologies, and clear communication about data handling practices were consistently recognised by participants as critical preconditions for their willingness to participate. This lesson reinforces the importance of making privacy-by-design choices visible and understandable to end users throughout the pilot, not only at the consent stage.
- **Real-time data processing and operational readiness must progress in parallel.** The current batch-processing approach to the ETL pipeline and model inference introduces a lag between data collection and the availability of actionable outputs for caregivers. Efforts to move the pipeline closer to real-time operation are already underway and are a priority for the replication phase. This lesson has direct implications for the operational usefulness of the system and for caregiver trust: timely, coherent information is more valuable and more likely to be acted upon than retrospective summaries.
- **Applying the FAITH AI_TAF methodology supports structured reflection and governance improvement.** The use of trustGuard, trustSense, and the AI Model Hub in the context of LSP7 has demonstrated that structured trustworthiness assessment activities are feasible within an active ageing pilot and provide practical value beyond compliance. In particular, the structured identification of vulnerabilities, the assessment of team maturity, and the definition of countermeasures have supported a more disciplined approach to AI governance and have helped the team prioritise improvement actions in areas such as privacy readiness, documentation practices, and resilience planning.

# 3 Application of FAITH Tools

## 3.1 Application of the FAITH AI_TAF V1 methodology

Section 3.1 presents the application of the FAITH AI Trustworthiness Assessment Framework (FAITH AI_TAF) v1 to the LSP7 active ageing pilot through a sequence of six structured phases: cartography, threat analysis, impact assessment, vulnerability assessment, risk analysis, and countermeasures. Together, these phases provide a systematic way to move from the definition of the AI system boundary to the identification of concrete mitigation actions, ensuring that trustworthiness is assessed not only in terms of model performance, but also in relation to the assets, people, and operational context involved in the deployment.

In LSP7, the cartography phase establishes the assessment scope by identifying the AI-enhanced Vitana solution, its core assets, their lifecycle stage, and the ownership context, while also characterising the system as medium criticality in the active ageing setting. The subsequent phases build directly on this foundation: threat analysis catalogues both adversarial and operational threats; impact assessment determines which trustworthiness dimensions would be most affected; vulnerability assessment examines the exposure of the declared assets; risk analysis combines likelihood, impact, and vulnerability into an overall risk picture; and countermeasures define the controls needed to reduce the most relevant risks and support trustworthy deployment.




*Figure 3 Phases of FAITH AI_TAF V1 methodology*

### 3.1.1 Cartography

Within LSP7, the cartography activity establishes the scope of the trustworthiness assessment by determining what exactly is being assessed, under which operational conditions, and through which asset boundaries. For this pilot, the system considered in the assessment is the AI-Enhanced Vitana solution in the Active Ageing domain, examined at a stage where core functionalities are still under development while some components are already closer to

deployment maturity. Rather than starting from the wider socio-technical environment, the assessment first isolates the relevant system elements so that each one can be described, documented, and later evaluated in a controlled and traceable manner. This delimitation is important because the first version of the FAITH AI_TAF assesses assets individually before moving toward broader system-level interdependencies.

The mapping exercise carried out for LSP7 identifies both model and data assets that define the current system boundary. The model layer includes M1, the clustering and segmentation model, and M2, the anomaly detection model, while the data layer includes DA1, the Activage DS VLC dataset, DA2, the VITANA dataset, and DAx, the FAITH LSP7 generated dataset. These assets do not all share the same maturity level, which is itself relevant for the assessment: M1 and M2 are positioned in the develop stage, DA1 is already in deploy, DA2 remains in develop, and DAx is in design. The cartography therefore captures not only the list of assets, but also their position in the lifecycle and their ownership context, which in this case is associated mainly with the ACTIVAGE and ACTIVAGE/BRIDG teams.

This phase also records contextual elements needed for the later assessment steps, especially the relationship between asset ownership, team maturity, and intended use. In the LSP7 questionnaire, the teams linked to the declared assets are reported with a High AI maturity level, which is relevant because the FAITH methodology considers organizational preparedness as part of the broader trustworthiness picture. The system is aimed at supporting monitoring and interpretation tasks in an active ageing setting, where outputs may influence care-related decisions by relatives or caregivers, even though the solution is not framed as an emergency or essential-service system. This makes the cartography phase particularly important, since it anchors the later analysis in the practical realities of the pilot rather than in an abstract description of the technology.

A further outcome of the cartography step in LSP7 concerns the characterization of system criticality. According to the information captured in the assessment materials, the AI system is not classified as a high-risk AI system under Article 6 and Annex III of the AI Act[4], and the deploying entity is not treated as an essential or important entity under the relevant NIS2[5] criteria referenced in the questionnaire. At the same time, the system is not treated as low-stakes, since it operates in the broader health and social care space and interacts with the well-being of older adults. For this reason, the assessment materials indicate a medium criticality interpretation, reflecting the fact that the system may affect users' safety, confidence, and care quality, but does not carry the same level of direct clinical or emergency consequence as highly regulated medical AI systems.

The cartography for LSP7 establishes a structured baseline for the subsequent phases of FAITH AI_TAF. This phase involves identifying the AI system, specifying its assets, documenting the lifecycle stages and ownership, and clarifying the criticality level of the pilot context. This

[4] https://artificialintelligenceact.eu/
[5] https://eur-lex.europa.eu/eli/dir/2022/2555/oj

foundational work creates the necessary framework for conducting threat analysis, impact assessment, vulnerability evaluation, and risk estimation.

In the specific case of LSP7, it underscores a significant characteristic of the pilot: the presence of assets at varying maturity levels within a single active aging solution. This diversity necessitates careful boundary definitions to ensure a coherent trustworthiness assessment.

### 3.1.2 Threat Analysis

In LSP7, the threat analysis phase focused on the structured identification of events and conditions that could negatively affect the trustworthiness, security, reliability, or proper functioning of the AI-Enhanced Vitana system and its associated assets. Following the FAITH AI_TAF v1 logic, the analysis was carried out at asset level, meaning that potential threats were considered in relation to the declared models and datasets rather than to the full socio-technical ecosystem in which the pilot may eventually operate. This approach made it possible to examine the exposure of the main technical components in a consistent way, while preserving traceability between each identified threat and the specific asset, lifecycle stage, and operational assumptions recorded during cartography.

The threat identification process considered both malicious and non-malicious sources of harm. On the one hand, the assessment included AI- and cybersecurity-related threats such as poisoning, evasion, model disclosure, data disclosure, and compromise of ML application components. On the other hand, it also covered operational and organisational threats, including human error, system malfunction, and failures related to insufficient protection, weak access management, or incomplete incident handling practices. This broader framing is important in the active ageing context, where the trustworthiness of the system depends not only on resistance to deliberate attacks, but also on the robustness of data pipelines, model behaviour, maintenance practices, and human interaction with the system outputs.

For LSP7, the assessment of threat frequency was informed by the characteristics of the pilot and by the more general experience from healthcare-related AI environments, as explicitly noted in the questionnaire. Several adversarial threats were judged to be possible but relatively infrequent. For example, evasion attacks were assessed as occurring approximately once every five years, while oracle attacks and model disclosure were estimated as less frequent, around once every ten years. By contrast, some operational threats were considered more relevant in practice, especially failure or malfunction of the ML application, which was assessed at once a year, reflecting the realistic possibility of disruptions caused by data drift, sensor malfunction, unexpected user behaviour, or service-level limitations in the operational environment.

The analysis also highlighted that some threats were considered unlikely under the current pilot conditions. Label modification and poisoning were both assessed as never in the questionnaire, mainly because the training and data management processes were understood as taking place in a controlled environment where external manipulation of the data pipeline is limited. Similarly, attacks requiring privileged technical visibility, such as adversarial examples crafted under white-box or grey-box conditions, were also judged to be rare in practical deployment settings. At the same time, the assessment did not dismiss disclosure-related threats, since both data disclosure

and model or data disclosure were still recognised as credible medium-term concerns, especially in relation to access control weaknesses, documentation exposure, third-party interactions, or insufficiently restricted outputs.

An important outcome of this phase was the recognition that the most relevant threat landscape for LSP7 is not dominated by highly sophisticated attacks alone. Instead, the assessment suggests a more balanced profile in which moderate-frequency operational failures and governance weaknesses may be at least as relevant as classical adversarial ML attacks. Threats such as compromise of ML application components and human error were assessed at around once every two years, indicating that software dependencies, configuration issues, user misunderstandings, or incomplete security integration may represent meaningful sources of risk for the pilot. This is especially relevant for an active ageing solution such as AI-Enhanced Vitana, where the practical value of the system depends on stable performance, appropriate handling of behavioural and sensor data, and sustained confidence from caregivers and other users.

The threat analysis phase in LSP7 resulted in a structured threat catalog that aligns with the declared assets and lifecycle stages of the pilot. The analysis reveals that the system operates in a relatively controlled environment, with a low expected frequency of severe adversarial manipulation. However, there are more significant concerns regarding malfunctions, data disclosure, component compromises, and vulnerabilities related to human factors. This assessment serves as the foundation for the subsequent FAITH AI_TAF phases, where the identified threats will be further analyzed in terms of their business consequences, vulnerabilities, and overall risk levels within the context of the active aging pilot.

### 3.1.3 Impact Assessment

In LSP7, the impact assessment phase examined the potential consequences that the previously identified threats could generate if they were to materialise during the development or operation of the AI-Enhanced Vitana system. In line with the FAITH AI_TAF v1 methodology, this phase did not focus only on technical failure as such, but on the broader business, legal, operational, and trust-related implications associated with the degradation of key trustworthiness properties. The assessment was therefore structured around the trustworthiness dimensions captured in the questionnaire, allowing the pilot team to determine which types of failure would have the most significant effect in the Active Ageing context.

The results indicate a differentiated impact profile rather than a uniformly high-consequence system. Some dimensions were assessed as having high business consequence, particularly privacy and interpretability. Privacy was rated high because a compromise could lead to significant legal and reputational consequences, including non-compliance with data protection obligations, loss of stakeholder trust, and possible sanctions linked to the processing of sensitive personal data. Interpretability was also considered high, since caregivers or other users may rely on the system's outputs to understand behavioural patterns or possible anomalies, and misinterpretation of these outputs could lead to inappropriate reactions, overlooked warning signs, or reduced confidence in the solution.

A second group of dimensions was assessed at a moderate consequence level. This includes security and resilience, fairness, accuracy, and robustness. For security and resilience, the assessment recognised that a compromised or unstable system could interrupt monitoring functions, reduce data quality, or delay the identification of relevant behavioural changes, thereby undermining continuity and trust in the service. Fairness was also rated moderate, reflecting the possibility that biased model behaviour could affect some user groups differently, with resulting reputational and governance consequences. In a similar way, insufficient accuracy or robustness could generate false positives, false negatives, or inconsistent outputs under changing behavioural or environmental conditions, which would not normally create immediate critical harm but could still weaken the usefulness and credibility of the system over time.

By contrast, some dimensions were classified with a low consequence level, namely safety, explainability, and transparency. This lower rating does not imply that these dimensions are irrelevant, but rather that the current LSP7 system is positioned as a supportive tool rather than as an emergency-response or essential clinical system. In the questionnaire, the impact of compromised safety was described as limited harm such as false alarms, missed detections, reputational effects, and loss of trust, while transparency and explainability were seen mainly as factors that influence stakeholder confidence, regulatory clarity, and user comfort rather than as immediate drivers of severe damage. This assessment is consistent with the broader characterization of the pilot as medium-criticality rather than high-risk under the AI Act or NIS2 criteria.

An important finding of this phase is that, in LSP7, the highest-impact concerns are closely linked to the way the system's outputs are understood and governed rather than to direct physical harm alone. In particular, the emphasis placed on privacy and interpretability shows that the most serious consequences in the active ageing setting may arise when users cannot trust the handling of sensitive data or cannot correctly understand the meaning of model outputs. This reflects the nature of AI-Enhanced Vitana as a system intended to support observation and decision-making around older adults' well-being, where the reliability of interpretation and the legitimacy of data use are central to acceptance and responsible deployment.

The impact assessment for LSP7 indicates a varied consequence landscape. The pilot demonstrates a high impact in areas such as privacy and interpretability, a moderate impact in resilience, fairness, accuracy, and robustness, and a lower impact in dimensions indirectly related to harm. This assessment serves as a crucial link between identifying threats and conducting subsequent risk analysis. It highlights that even infrequent threats should be prioritized if they affect areas with significant legal, ethical, or operational implications. Consequently, the impact assessment positions the LSP7 risk profile as one influenced more by the cumulative effects of trust, governance, and decision-support failures within an active aging context rather than by immediate, extreme dangers.

### 3.1.4 Vulnerability Assessment

In LSP7, the vulnerability assessment phase examined the extent to which the declared assets of the AI-Enhanced Vitana system remain exposed to the threats identified in the previous phase due to missing, partial, or insufficiently formalised safeguards. In accordance with the FAITH

AI_TAF v1 approach, this assessment was carried out through structured control questions linked to the system lifecycle stage and to the specific AI assets declared during cartography, namely M1, M2, DA1, DA2, and DAx. This phase therefore moved beyond identifying what could go wrong and focused instead on understanding where the current technical, organisational, and governance arrangements are still weak or incomplete.

The results show that LSP7 has a mixed vulnerability profile, with some controls already in place and others not yet implemented or only partially addressed. On the positive side, the assessment indicates that reliable data sources are used for the main datasets, role-based access control has been applied to the relevant data assets, both models are considered compliant with data security requirements, and both M1 and M2 are reported as compliant with identity management, authentication, and access control policies. Additional strengths include the fact that all data used by the models is reported as controlled, with the team explicitly noting ownership of the sensing infrastructure and therefore control over the data pipeline. Disclosure-related safeguards are also partially present, since both models are reported to provide only the necessary information in their outputs and to respect differential privacy according to the assessment responses.

At the same time, the questionnaire highlights several areas where vulnerability remains significant. One of the clearest gaps concerns adversarial robustness and poisoning awareness. Although M1 is described as using a resilient design and a more diverse training dataset, the exposure of both M1 and M2 to poisoning or adversarial attacks has not been formally assessed, and poisoning controls after model evaluation are reported as absent for both models. In addition, M2 appears weaker than M1 in this area, since it is explicitly marked as not having a resilient design and as lacking defined indicators of proper functioning. This suggests that, while some resilience thinking is present for selected components, adversarial risk management is not yet embedded consistently across the full model set.

A second cluster of vulnerabilities relates to security lifecycle management and ML-specific governance. The responses indicate that protections for test environments are not ensured for either model, vulnerabilities in the software components used by the models have not been systematically checked, and ML-specific elements have not been integrated into existing security policies. In addition, compliance with the relevant security policies is marked as NA for both models, with an explicit comment asking where those policies can be found, which suggests that the governance framework may not yet be sufficiently defined or operationalised for the AI assets under assessment. Likewise, inclusion of the models in broader asset management processes is also marked as NA, pointing to a partial disconnect between AI development practices and wider organisational security and governance structures.

Further vulnerabilities emerge in relation to privacy-enhancing architecture and advanced protective mechanisms. While the models are reported to limit output information and to meet differential privacy expectations, neither M1 nor M2 uses federated learning to reduce data breach risks, and neither model has restricted the available information about its architecture. These findings do not necessarily indicate critical weaknesses in the current pilot setting, which remains relatively controlled, but they do suggest that the system would require stronger

safeguards if deployed in a broader or more exposed operational environment. Similarly, the lack of integration of ML-specific cybersecurity awareness strategies for stakeholders indicates that human and organisational preparedness for AI-specific threats is still limited.

The vulnerability assessment for LSP7 indicates that the system has a solid foundation in data governance, access management, and basic output restrictions. However, it also reveals significant weaknesses in areas such as advanced robustness, continuous security integration, protection of test environments, and organisational governance regarding machine learning-specific risks. This assessment suggests that the system is not uncontrolled or immature, but rather operates like a pilot program. Its core controls are stronger in traditional data and access management compared to proactive AI security and lifecycle resilience practices. These findings align with previous threat and impact assessments and provide a crucial foundation for the next phase of the FAITH AI_TAF, in which likelihood, consequence, and vulnerability will be combined to derive the overall risk profile for the LSP7 pilot.

### 3.1.5 Risk Analysis

The risk analysis phase for LSP7 focused on the structured consolidation of the findings generated during the preceding threat analysis, impact assessment, and vulnerability assessment phases. In line with the FAITH AI_TAF v1 methodology, the objective of this step was to determine the relative significance of the identified risks affecting the AI-Enhanced Vitana system by considering, in combination, the estimated frequency of relevant threats, the severity of their potential consequences, and the degree of exposure revealed through the vulnerability assessment. This integrated view makes it possible to distinguish between lower-priority issues that mainly require monitoring and those risk areas that should be prioritised for mitigation within the pilot context.

For LSP7, the resulting risk profile can be characterised as moderate overall, but unevenly distributed across trustworthiness dimensions. This pattern is largely explained by the combination of a medium criticality level, relatively low frequency estimates for several advanced ML-specific attacks, and a set of business consequence ratings that remain significant in selected areas, particularly privacy and interpretability. More specifically, the questionnaire indicates low business consequence for safety, explainability, and transparency; moderate consequence for security and resilience, fairness, and validity-related aspects such as accuracy and robustness; and high consequence for lack of interpretability and privacy compromise. As a result, the highest-priority risks are not necessarily those associated with the most frequent threats, but rather those where moderate vulnerabilities intersect with dimensions that carry stronger organisational, legal, and trust-related consequences.

A first important cluster of risks concerns privacy, disclosure, and information governance. Although the frequency assigned to data disclosure and model-or-data disclosure is relatively low, the associated business consequence of privacy compromise is explicitly rated as high, reflecting the legal sensitivity of personal data and the importance of preserving user and stakeholder trust in an active ageing setting. This means that even infrequent disclosure scenarios remain materially relevant, especially where the assessment also shows incomplete controls regarding federated learning, limitation of architectural information, and the broader

integration of ML-specific governance measures into security policies and asset management processes. In practice, these risks should be considered priority areas because their likelihood may be constrained by the pilot environment, but their impact would be disproportionately significant if realised.

A second cluster relates to model reliability, interpretability, and operational continuity. The threat analysis suggests that failure or malfunction of ML applications is one of the more plausible operational risks in the broader sectoral context, with failures linked to data drift, sensor malfunction, or unforeseen behaviour estimated at approximately once a year. This becomes particularly important in LSP7 because the business consequence of insufficient interpretability is rated as high, while security and resilience, fairness, accuracy, and robustness are all assessed at moderate consequence levels. When these impact ratings are read together with the vulnerability profile, especially the lack of defined indicators of proper functioning for M2, the absence of poisoning controls after model evaluation, and the limited protection of test environments, the analysis points to a meaningful risk concentration around degraded model behaviour, misunderstood outputs, and delayed detection of system weaknesses.

From an adversarial and cybersecurity perspective, the overall risk appears contained but not negligible. Several attack types, such as poisoning, white-box adversarial examples, and model disclosure, are considered infrequent or rare in comparable environments, which reduces their immediate priority in the current pilot setup. However, the vulnerability assessment also shows that exposure to poisoning or adversarial attacks has not been formally assessed for either model, software component vulnerabilities have not been systematically checked, and ML-specific awareness and security-policy integration remain incomplete. This creates a familiar risk asymmetry: the most advanced attacks may be unlikely in the present environment, but the system is not yet sufficiently hardened against them should exposure increase in a future, more connected deployment scenario.

The LSP7 risk analysis reveals that the system does not exhibit uniformly high risks; instead, it identifies a specific set of medium to elevated risks concentrated in areas where failures in trustworthiness could have significant consequences for users, caregivers, compliance, and service credibility. The key priorities that emerge from this analysis are related to privacy protection, the interpretability of outputs, the reliability of model behaviour, and the enhancement of security and governance practices specific to machine learning. These findings will serve as the foundation for the next phase of the FAITH AI_TAF project, where concrete countermeasures will be developed to mitigate exposure in the most critical risk areas while ensuring the practical feasibility of the LSP7 pilot deployment.

### 3.1.6 Countermeasures

Following the identification and prioritisation of risks, the countermeasures phase for LSP7 focused on defining appropriate actions to reduce the likelihood, impact, and exposure associated with the most relevant trustworthiness risks affecting the AI-Enhanced Vitana system. In line with the FAITH AI_TAF v1 methodology, the purpose of this phase was not only to list generic controls, but to identify realistic and proportionate measures that respond directly to the weaknesses observed across the assessed assets, namely M1, M2, DA1, DA2, and DAx. Given the

pilot-stage nature of the system and its medium-criticality profile, the selected countermeasures should be understood as a combination of immediate reinforcement actions and medium-term improvements intended to strengthen the maturity of the AI system as it evolves.

A first group of countermeasures concerns robustness, malfunction detection, and operational reliability. The assessment indicates that M2 is the more exposed model in this respect, since it lacks both a resilient design declaration and defined indicators of proper functioning, while poisoning controls after model evaluation are absent for both models. For this reason, priority actions should include the definition of explicit health and performance indicators for M2, the extension of post-evaluation monitoring practices to both models, and the establishment of simple but regular checks for drift, anomalous outputs, and unexpected behaviour under realistic use conditions. These actions would help reduce the risk of undetected degradation, support more reliable interpretation of system outputs, and improve the ability of the technical team to intervene before problems propagate into service use.

A second set of countermeasures should address adversarial and ML-specific security weaknesses. Although the current pilot operates in a relatively controlled context and several advanced attacks are assessed as infrequent, the vulnerability analysis shows that the exposure of M1 and M2 to poisoning or adversarial attacks has not been formally assessed, and that ML-specific awareness measures have not yet been integrated into the existing security approach. In response, a proportionate mitigation strategy would include a lightweight exposure assessment for both models, the introduction of documented checks for adversarial and poisoning-related vulnerabilities during model review, and the inclusion of AI-specific threat awareness in internal team practices. For LSP7, this does not necessarily require complex defence architectures at this stage, but it does require moving from implicit assumptions of safety toward explicitly documented robustness and security practices.

A third priority area relates to security lifecycle management and governance integration. The assessment shows that test-environment protections are not ensured, component vulnerabilities are not systematically checked, ML-specific security requirements are not integrated into existing security policies, and the inclusion of AI assets in broader asset management processes is still unclear or not established. Accordingly, an important countermeasure package for LSP7 would be to formalise a minimum governance baseline for AI assets, including secure handling of development and test environments, periodic review of software dependencies and components, clarification of applicable security policies, and explicit registration of the models within organisational asset-management and security processes. These actions are particularly important because several weaknesses identified in the assessment do not arise from missing model functionality alone, but from the partial disconnect between ML development practices and the wider organisational governance framework.

Additional countermeasures are needed in the area of privacy and information disclosure control. While the assessment indicates that both models provide only the necessary output information and are reported to respect differential privacy, it also shows that federated learning is not used and that information about the model architecture has not been intentionally limited. In the current pilot setting, these gaps do not automatically imply severe weakness, but they

suggest useful reinforcement measures for future scalability. Suitable countermeasures would therefore include reviewing what technical and descriptive information about the models is externally visible, documenting clear rules on output minimisation and access to model internals, and evaluating whether more advanced privacy-preserving techniques would become necessary in future distributed or multi-site deployments. This would help preserve the currently strong privacy posture while preparing the system for broader operational exposure.

The countermeasure strategy for LSP7 should be viewed as a phased strengthening process rather than a complete redesign of the system. The system already demonstrates important strengths in data-source reliability, role-based access control, data security, and basic access management practices, which provide a solid foundation for further improvement.

The key next steps are to enhance areas where the current pilot is relatively weaker, specifically in monitoring model behavior, assessing exposure to adversarial threats, protecting the test environment, establishing ML-specific governance, and formally integrating AI assets into security and asset management processes.

In this context, the countermeasures phase translates the insights from the earlier FAITH AI_TAF steps into a practical roadmap for enhancing the trustworthiness and resilience of the LSP7 pilot over time.

## 3.2 Tool Validation Planning & Training Processes

### 3.2.1 Dry-run

Before the formal launch of the FAITH tool validation activities, a dry-run was organised within the Active Ageing LSP to check the readiness of the assessment process, familiarise the participants with the tools, and identify practical issues that needed to be resolved before full deployment. The exercise mainly concentrated on the preliminary use of the FAITH risk assessment questionnaire and the preparatory workflows associated with trustGuard, in the context of the Vitana-based home monitoring solution.

The dry-run involved representatives from the LSP7 consortium, mainly including technical staff, who jointly reviewed the structure, wording, and applicability of the questionnaire to the active ageing use case. Particular attention was paid to ensuring that the assessment items could be interpreted consistently for a privacy-preserving system that monitors behavioural patterns through non-intrusive sensors rather than cameras or microphones. The team also verified that the questionnaire covered the main elements relevant to the pilot, such as the AI assets, the operational context, the intended users, and the trustworthiness characteristics linked to elderly care.

During the dry-run, the participants simulated the complete assessment workflow using the LSP7 AI services as reference assets. This included a first pass at defining the system boundaries, identifying the lifecycle stage of the AI components, and reviewing the most relevant trustworthiness dimensions for the pilot, such as privacy, explainability, robustness, fairness, and security. The group also tested how the information collected during the exercise would support

later stages of the methodology, especially threat analysis, vulnerability assessment, risk estimation, and mitigation planning.

The dry-run further served as a calibration exercise among the participants. Through guided discussion, the team aligned their understanding of key assessment concepts, including likelihood, impact, and criticality, which helped to improve consistency for the formal validation phase. In addition, the group identified terminology that required adaptation to the active ageing setting, especially where general AI-risk concepts had to be translated into the realities of home-based monitoring, caregiver interaction, and the use of behavioural indicators for early warning purposes.

From an operational perspective, the dry-run confirmed that the selected assets were appropriate for assessment and that the involved teams were sufficiently prepared to proceed. It also highlighted the need for concise supporting evidence related to the deployment architecture, sensor configuration, model documentation, consent procedures, and user interaction pathways. As a result, several refinements were introduced in the preparatory materials and internal guidance used for the subsequent trustGuard assessment and the wider FAITH tool validation process.

### 3.2.2 Plan development

The Plan Development for LSP7 – Active Ageing aimed to establish a structured implementation roadmap to support the deployment, validation and iterative improvement of FAITH technologies within the Active Ageing LSP. It focused on translating high-level pilot objectives into a practical and operational implementation plan, which results as a living implementation framework covering deployment activities and supporting the transition from preparatory work towards real operational settings. To ensure operational execution, the LSP7 implementation plan included a timeline with assigned responsibilities and target periods for each activity, structured around project months and defined ownership among technical partners, pilot representatives and risk assessment stakeholders.

***Planning activities***

Initial planning activities focused on defining and validating the implementation strategy for LSP7. This stage included the formal definition of the implementation roadmap and agreement among participating partners regarding priorities, sequencing of activities and deployment objectives. The planning process ensured a common understanding of responsibilities and created a baseline for monitoring progress throughout implementation.

*Table 1. Plan development · planning activities*

| Task | Assigned to | Start | End |
|---|---|---|---|
| Define plan | AOA & BRIDG | M25 | M25 |
| Agree plan | AOA & BRIDG | M25 | M25 |

***Training and support activities***

Training and support activities were incorporated as key prerequisites for deployment readiness. The implementation plan identified several capacity-building actions intended to prepare partners and users for interaction with FAITH tools and infrastructures. These activities covered training on AI Model Passport functionalities and AI Data Passport components, together with onboarding activities for trustSense and trustGUARD tools. In addition, the plan incorporated SSM-related training and introductory sessions focused on FAITH STM platform registration and operational workflows. These activities aimed to ensure that all participants acquired the necessary competencies and technical understanding required before entering broader implementation stages, also ensure sufficient technical preparedness before wider deployment activities.

*Table 2. Plan development · training and support activities*

| Task | Assigned to | Start | End |
|---|---|---|---|
| Attend training for AI Model Passport (pythonic library) | AOA & BRIDG | M25 | M26 |
| Attend training for AI Data Passport | AOA & BRIDG | M25 | M26 |
| Attend training for trustSense | AOA & BRIDG | M23 | M23 |
| Attend training for trustGUARD | AOA & BRIDG | M26 | M27 |
| Attend training for SSM | AOA & BRIDG | M32 | M33 |
| Attend introduction to FAITH STM Platform registration and workflows | AOA & BRIDG | M33 | M33 |
| Attend training for AI Model Passport (pythonic library) | AOA & BRIDG | M25 | M26 |
| Attend training for AI Data Passport | AOA & BRIDG | M25 | M26 |

***FAITH tools deployment and iterative implementation***

The central component of the implementation plan concerned the deployment and iterative refinement of FAITH technologies within LSP7. The roadmap considered deployment in both SANDBOX and real environments. The implementation strategy followed a progressive approach allowing technologies to be introduced, assessed and improved over time. The implementation activities included deployment of the System Under Test (SUT) within SANDBOX environments followed by progressive introduction into real operational settings. Alongside deployment, the process incorporated team maturity assessment activities through trustSense, together with structured collection of usability and usefulness feedback. Results from these assessments supported the identification and implementation of mitigation measures, creating an iterative cycle in which observations and outcomes continuously informed subsequent improvements. This iterative process was intended to support continuous adaptation of the deployed solutions and ensure alignment with user expectations and operational requirements.

*Table 3. Plan development · FAITH tools deployment and iterative implementation*

| Task | Assigned to | Start | End |
|---|---|---|---|
| Deploy SUT in SANDBOX | BRIDG / AOA | M18 | M36 |
| Deploy SUT in REAL ENVIRONMENT | BRIDG / AOA | M22 | M40 |
| Core Model replication for cooking control (TBC) | BRIDG | M24 | M36 |
| Core Model replication for sedentarism control (TBC) | BRIDG | M24 | M36 |
| Core Model replication for bathroom usage (TBC) | BRIDG | M24 | M36 |
| Deploy SUT in Replication Pilot | BRIDG / AOA | M36 | M38 |

***Trust assessment and monitoring activities***

The implementation roadmap integrated trust-related assessment activities from the early deployment stages. Specific tools and methodologies were incorporated to monitor maturity, usability and trustworthiness dimensions. Assessment activities included the use of trustSense and trustGUARD tools to support continuous evaluation and identify areas requiring corrective actions. Feedback generated through these mechanisms was incorporated into subsequent implementation cycles through mitigation and refinement activities.

*Table 4. Plan development · Trust assessment and monitoring activities*

| Task | Assigned to | Start | End |
|---|---|---|---|
| Use trustSense for team maturity assessment (early deployment period) | Risk Assessor + End Users | M23 | M25 |
| Provide formal feedback on Usability and Usefulness of trustSense | Risk Assessor | M25 | M25 |
| Implement trustSense mitigations | AOA / BRIDG | M25 | M36 |
| Use trustGUARD v1 for in-isolation AI Model Asset assessment @ deployment phase | Risk Assessor | M27 | M28 |
| Use trustGUARD v1 for in-isolation (Input) Data Asset assessment @ deployment phase | Risk Assessor | M27 | M28 |
| Provide formal standardised feedback on Usability and Usefulness of trustGUARD v1 during deployment phase | Risk Assessor | M28 | M28 |
| Apply 1 or more Controls to AI Model Asset or Data Asset during deployment phase based on trustGUARD recommendations | BRIDG / AOA | M29 | M32 |
| Update and re-run in-isolation AI Model Asset and/or Data Asset assessment including new Control(s) using trustGUARD v1 | Risk Assessor | M30 | M32 |
| On-going iteration of trustGUARD v2 based on LSP feedback (including on-going Knowledgebase updates) | ATC | M30 | M36 |
| Use SSM for in-context SUT assessment | Risk Assessor | M33 | M35 |

| Task | Assigned to | Start | End |
|---|---|---|---|
| Provide formal standardised feedback on Usability and Usefulness of SSM | Risk Assessor | M35 | M35 |
| On-going iteration of SSM v2 based on LSP feedback (including on-going Knowledgebase updates) | UOS | M36 | M45 |

***AI Model Hub and platform integration activities***

Additional activities within the implementation plan included the creation and management of AI assets and associated platform components. These tasks supported integration with broader FAITH ecosystem services and ensured long-term sustainability of developed solutions.

*Table 5. Plan development · AI Model Hub and platform integration activities*

| Task | Assigned to | Start | End |
|---|---|---|---|
| Create asset entries in AI Model Hub | Juan Carlos Naranjo, BRIDG | M26 | M28 |
| Provide formal standardised feedback on Usability and Usefulness of AI Model Hub | Juan Carlos Naranjo, BRIDG | M28 | M28 |
| Iterate AI Model Hub v2 based on LSP feedback | FORTH / CNR | M26 | M32 |
| Use FAITH STM Platform | Risk Assessor | M30 | M33 |
| Provide formal standardised feedback on Usability and Usefulness of FAITH STM Platform | Risk Assessor | M33 | M33 |
| Iterate FAITH STM Platform | AOA / BRIDG / UOS | M34 | M36 |
| Create SSM system model | AOA / BRIDG / UOS | M33 | M33 |

### 3.2.3 Tool Training

The tool training activities for the Active Ageing LSP were organised as targeted sessions for the technical AI teams of AOA and BRIDG, with the objective of preparing the partners for the structured use of the FAITH tools in the context of LSP7. Attendance was intentionally restricted to the technical AI team in order to ensure a focused and operationally relevant transfer of knowledge.

The sessions were held on 10/11/2025, 14/01/2026 and 25/02/2026 and covered the tools and workflows considered most relevant for the first trustworthiness assessment activities in the Active Ageing domain. The participants were introduced to the scope, logic, and expected outputs of the FAITH tools, with emphasis on their application to the Vitana-based home monitoring solution and its associated AI components, including behavioural segmentation and anomaly detection services. Particular attention was given to how the tools could support the documentation, assessment, and iterative refinement of AI assets operating in a privacy-preserving home environment.

Training on the AI Model Passport and Data Passport focused on the structured documentation of datasets, models, and related metadata within the FAITH ecosystem. The technical teams were guided through the registration of AI assets, completion of descriptive and governance-related metadata, linkage between models and datasets, and the documentation of lifecycle information relevant to the LSP7 solution. This was especially important for the Active Ageing pilot, where traceability of the data sources, model configurations, and deployment context is necessary to support transparency, reproducibility, and subsequent trustworthiness assessment activities.

Training on trustSense was limited to the perspective of the technical AI team and concentrated on the interpretation and completion of the maturity assessment process relevant to developers and technical maintainers of the system. The sessions explained the rationale of the questionnaire, the dimensions being assessed, and the way in which the results can inform improvement actions related to team readiness, governance maturity, and trustworthy AI development practices. This focus is consistent with the LSP7 validation approach, where the technical teams of AOA and BRIDG play a central role in the deployment, maintenance, and continuous refinement of the AI-enabled monitoring solution.

Training on trustGuard addressed the practical execution of the FAITH AI_TAF v1 assessment workflow, including asset identification, threat analysis, impact estimation, vulnerability assessment, risk calculation, and the definition of mitigation measures. During these sessions, the participants were guided on how to apply the methodology to the specific characteristics of the Active Ageing use case, taking into account the cloud-based infrastructure, the use of non-intrusive sensors, the handling of behavioural data, and the interaction between AI outputs and caregiver-facing applications. The training also clarified the role of supporting evidence and internal documentation needed to complete the assessment in a consistent and traceable way.

In addition to the formal sessions, follow-up exchanges were carried out between AOA and BRIDG technical members to consolidate understanding of responsibilities, expected deliverables, and sequencing of the validation tasks. These exchanges helped align the two partners on the practical use of the tools and facilitated a shared interpretation of the trustworthiness dimensions most relevant to LSP7, notably privacy, explainability, robustness, security, and fairness. As a result, the technical AI team was adequately prepared to proceed with the execution of the FAITH tool validation activities within the first pilot phase of the Active Ageing LSP.

## 3.3 FAITH Tool Validation Activities

### 3.3.1 trustSense

trustSense was employed in the context of LSP7 to capture how different participant profiles involved in the pilot perceived and enacted trustworthiness-related practices during the use and development of the FAITH solutions. In practical terms, the tool was used as a structured questionnaire-based assessment instrument to gather evidence on behavioural, organisational, and operational aspects linked to trustworthy AI adoption, rather than to evaluate the AI artefact alone.

For LSP7, the roles concern the **AI Technical Team / AI developer profile**, which represents the group responsible for configuring, adapting, maintaining, and technically supporting the pilot solution. The assessment output reports an overall trustworthiness level of **Moderate** with a calculated score of **3.42**, indicating that the team demonstrates several positive trustworthiness-enabling practices, but still requires further consolidation in a number of dimensions before reaching a consistently high maturity level.

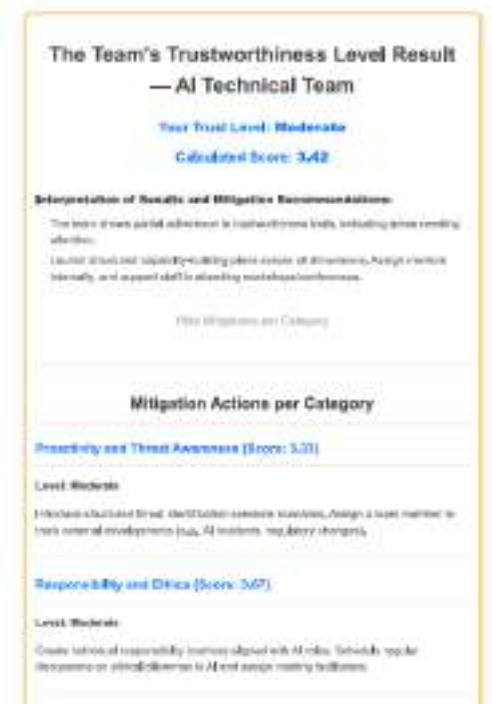



*Figure 4 trustSense AI technical team result*

The detailed category results help explain this overall picture. The strongest areas are **Collaboration**, **Integrity**, and **Technical Proficiency**, each scoring **4.00**, which suggests that the technical users are effective in knowledge sharing, show solid professional conduct, and possess the practical skills needed to support AI validation and problem detection in the pilot setting. This reading is also aligned with the questionnaire responses, where the team strongly recognised the intended use and expected outcomes of the AI systems, and showed positive ratings in data quality management, collaboration, advanced technical tooling, and ethical conduct.

At the same time, the assessment highlights a set of dimensions that remain only moderately developed. **Proactivity and Threat Awareness** scored **3.33**, while **Responsibility and Ethics** scored **3.67**, and **Problem Solving**, **Resource Accessibility**, **Policy Adherence**, and **Motivation and Commitment** each scored **3.00**. These results indicate that, although the team has a functional baseline for trustworthy AI work, a more systematic approach is still needed in areas such as structured threat anticipation, clearer role allocation, reinforcement of internal policy uptake, and more formalised continuous-improvement routines.

Two dimensions stand out as the clearest improvement priorities in LSP7. **Innovation and Adaptability** and **Resilience** both scored **2.50**, showing that the team would benefit from more

explicit mechanisms for introducing trustworthiness-related improvements, documenting recovery procedures, and embedding resilience checkpoints into project execution. Most notably, **Privacy and Compliance** received the lowest score, **2.00**, and the mitigation guidance explicitly recommends foundational training on personal data handling together with simplified privacy-by-design guidance, making this the most critical gap identified for the technical user type.

From a tool-validation perspective, the use of trustSense in LSP7 proved useful because it translated broad trustworthiness principles into a profile-specific maturity view that can support targeted follow-up actions. For the **technical/developer user type**, the conclusion is that LSP7 benefits from a capable and collaborative team with sound technical competence, but one that still needs stronger routines in privacy compliance, resilience planning, and structured innovation in order to support trustworthy deployment more robustly over time.

Considering the results by **user type**, the clearest conclusion is therefore as follows: technical users already provide a reliable operational basis for the pilot, especially in collaboration, integrity, and engineering capability, but their maturity is uneven across governance-oriented dimensions. This means that, in LSP7, trustSense was not only used to describe the current state of the technical team, but also to reveal where organisational strengthening should be prioritised, particularly for privacy awareness, recovery preparedness, and more structured trustworthiness improvement practices.

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

**Data Passport**

The dataset registered within the AI Model Hub is a pre-existing dataset called LOCS. In the Model Hub it is registered as “Elderly people home activity for active aging from December 2018 to June 2024” and can be found here. The data from LOCS comes from a very similar setup to the one used during the pilot. The data is from a previous project of AOA and all the details can be found in the provided link.

After successfully registering the dataset, feedback about the use of the tool in LSP7 was submitted. The main issues encountered during the process revolved around the interface for the dataset registration. It asks for an extensive amount of metadata that not a single person has the answer to, spread across several web forms. With this format it is hard to coordinate and complete the registration process, which is further exacerbated by the existence of hidden specifications about the form fields, and an opaque error message when not all these specifications are satisfied.




*Figure 5. LSP7 Data Passport*

**AI Passport**

The current model from LSP7 consists of a clustering model for behavioral patterns in senior population at their homes. The registration to the AI hub was done through the AI Passport library maintained by FORTH, instead of the manual alternative as in the Data Passport. The library was installed, configured with the necessary metadata, and implemented within our training pipeline. The pipeline starting point is from an already processed set of data, which is precomputed outside of the pipeline itself due to the heavy computation involved in transforming the raw data into something that can be ingested by the clustering model. This data is derived directly from the LOCS dataset and explicitly linked through its unique Data Passport identifier in the Model Hub. The model is registered under the name “faith_lsp7_clustering_model” and can be found here. Close communication was maintained with a FORTH representative for troubleshooting the registration of both the model and the dataset, and making sure the final outcomes worked for both sides.




*Figure 6. LPS7 AI Model passport*

### 3.3.3 trustGuard

trustGuard was used in LSP7 as the structured mechanism for translating the trustworthiness assessment methodology into an operational review of the deployed AI asset. The tool was applied to the **LSP7 clustering model**, which was assessed in **production** and assigned **medium** criticality, so the workflow focused on understanding the model's exposure, the consequences of failures, and the safeguards already in place.

The assessment was organised as a stepwise exercise. First, the asset was defined and contextualised through cartography, including the model name, lifecycle stage, and criticality level. From there, the evaluation moved into the threat, consequence, and vulnerability sections, allowing the assessor to build a traceable risk picture for the system rather than relying on isolated judgments.

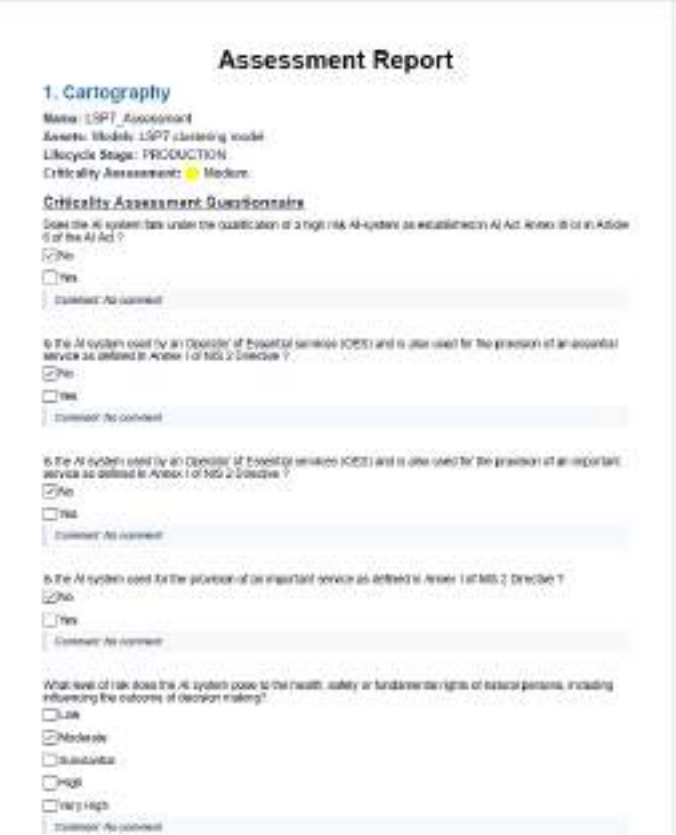



*Figure 7. Fist page LSP7 trustGuard Assesment*

In the threat analysis phase, trustGuard was used to review a broad catalogue of possible AI-related threats, covering areas such as evasion, poisoning, model or data disclosure, compromise of model components, human error, regulatory non-compliance, reduced interpretability, and bias-related behaviour. The value of the tool in this stage was that it converted a wide risk space into a repeatable questionnaire, making it easier to compare threats and keep the assessment consistent across dimensions.

The consequence assessment then connected each trustworthiness property to its possible operational impact. Instead of treating threats only as technical issues, trustGuard required the assessor to judge their potential business relevance in terms of safety, resilience, explainability, privacy, fairness, transparency, and robustness. This helped frame the evaluation around the real effect that a compromise could have on the use of the model in the LSP7 environment.

trustGuard was also used to examine whether relevant controls were already active. The vulnerability section checked for elements such as security monitoring, attack-surface awareness, access control, privacy-preserving techniques, data sanitisation, and lifecycle oversight. In practice, this stage served to map the current protective posture of the model and to show where the deployment was already covered and where additional safeguards were still missing.

The resulting risk output showed a largely contained profile. Most threats were scored as **Low**, including model or data disclosure, compromise of ML components, poisoning, reduced interpretability, privacy-related threats, and several bias-related issues. The most relevant exception in the extracted results was **Human error**, which was rated as **High** risk, indicating that the main exposure was not the clustering model alone but the way the surrounding human process could influence its safe and trustworthy use.

The mitigation stage was used to register concrete responses for the identified issues. For several low-risk items, the report noted that no mitigation action was required, while other cases were linked to controls such as monitoring proper functioning indicators, reducing model information exposure, strengthening security policy alignment, improving asset-management integration, and applying documentation requirements to AI projects. For the higher-risk human-error case, the suggested actions again pointed to documentation and process integration, confirming that procedural discipline was the priority area for intervention.

In LSP7, trustGuard served as a decision-support tool for tracing risks, prioritizing actions, and planning mitigations. It offered a structured approach to demonstrate that the model posed minimal direct risk in most categories. However, it also highlighted that the effectiveness of the deployment relied significantly on operational controls and human management.

For LSP7, the main conclusion is that trustGuard worked as a practical bridge between methodology and deployment reality. It helped identify where the system was already reasonably protected and where further work should focus, especially on reducing human-error exposure, tightening documentation, and formalising routine controls around the model's lifecycle.

## 3.4 Threat Analysis

The threat analysis carried out in LSP7 combined the evidence generated by the FAITH tools at asset level, looking both at the human side of the pilot and at the AI asset itself. In our case, this meant examining the maturity and readiness of the AI technical team through trustSense, alongside the threat profile of the LSP7 clustering model in production through trustGuard.

From the LSP7 perspective, this combined view was especially relevant because the threats identified during the assessment were not limited to technical failures of the model. Our assessment activities showed that part of the exposure also comes from the way the system is governed, maintained, monitored, and used by the people involved in the pilot.

For the human asset, the trustSense results for the AI technical team indicated an overall Moderate trustworthiness level, with a score of 3.42. In our interpretation, this means that the LSP7 team already has a solid basis to support trustworthy AI deployment, but still presents some weaknesses that may increase exposure to operational and organisational threats if they are not progressively addressed.

The areas that appeared most sensitive for our team were Privacy and Compliance with a score of 2.00, followed by Innovation and Adaptability and Resilience, both with 2.50. In practical terms, these findings suggest that, within LSP7, the most relevant human-side threats are linked

to insufficiently mature privacy practices, limited preparedness for incident recovery, and the absence of more systematic mechanisms to introduce mitigation improvements when new trustworthiness concerns emerge.

This reading is also supported by the detailed questionnaire responses. In our case, the team did not report strong agreement on the existence of established routines for introducing mitigation actions, showed only neutral confidence regarding recovery capacity after incidents, and gave a weak result on privacy and legal compliance prioritisation. These points are important because they indicate that the most critical issues in LSP7 are not related to lack of technical understanding, but rather to the need for stronger procedural discipline and more mature governance practices.

At the same time, the trustSense assessment also showed that the LSP7 technical team has clear strengths. Collaboration, Integrity, and Technical Proficiency all scored 4.00, while the detailed responses confirmed strong understanding of intended system use, good awareness of multidimensional trustworthiness requirements, and solid confidence in data-related and technical tasks. This is important in our pilot context, since it means that the main weaknesses do not come from insufficient expertise, but from uneven maturity across specific trustworthiness dimensions.

For the technical asset, trustGuard was applied to the LSP7 clustering model, assessed in production and classified with medium criticality. In our assessment process, the tool allowed us to review a broad set of asset-level threats, including evasion, poisoning, disclosure of model or data, compromise of ML components, malfunction of the application, human error, lack of interpretability, privacy issues, bias-related threats, drift, and accountability-related weaknesses.

The resulting picture for the model was relatively balanced. Most of the technical threats identified for the LSP7 clustering model were assessed as Low risk, including disclosure-related threats, poisoning, compromise of components, malfunction, reduced interpretability, privacy-related threats, and several fairness-related issues. From our point of view, this indicates that the model itself does not currently appear to concentrate severe asset-level technical risks across all trustworthiness properties.

However, one result stood out clearly in our LSP7 assessment: Human error was rated as High risk. This is a particularly relevant finding for the pilot, because it shows that the main source of exposure is not necessarily the algorithm alone, but the way the asset is operated, interpreted, and managed in practice. In other words, the most significant threat identified in LSP7 emerges at the intersection between the model and the surrounding human process.

When we consider the trustGuard and trustSense results together, the overall threat picture becomes more coherent. trustGuard identifies human error as the main explicit threat affecting the production asset, while trustSense shows that our technical team still has room for improvement in privacy compliance, resilience, mitigation planning, and structured threat awareness. Taken together, these findings suggest that, in LSP7, the most relevant threats are socio-technical rather than purely algorithmic.

The threat analysis in LSP7 indicates that the LSP7 clustering model has a relatively low direct technical risk profile. However, the main vulnerabilities are found in the organizational and operational aspects surrounding it. Based on our pilot activities, the most significant threats are related to gaps in privacy and compliance, a lack of resilience and procedural maturity within the technical team, and the high risk of human error in the usage and maintenance of the production asset.

## 3.5 Initial Risk Assessment results

The initial risk assessment for LSP7 was developed by combining the outputs of trustSense and trustGuard into a single early-stage view of risk exposure. Rather than revisiting each individual threat identified in the previous section, this step focused on interpreting the overall risk posture of the pilot and on understanding which factors are currently driving that posture.

At this stage, the overall picture for LSP7 can be described as moderate and uneven. The available evidence does not suggest a broadly critical risk condition across the assessed asset, but it does show that the current trustworthiness baseline still depends on a number of safeguards that are not yet fully formalised or systematically embedded in operational practice.

From the pilot perspective, one important element in this initial assessment is that the LSP7 clustering model was assessed in production with medium criticality. This means that the risk profile cannot be interpreted as negligible, even if many of the assessed technical risk outcomes remained at low level. Instead, the production setting requires attention to operational continuity, proper oversight, and the ability to react consistently if issues arise during real use.

The trustSense results reinforce this interpretation. The AI technical team obtained an overall trustworthiness score of 3.42, corresponding to a Moderate level, which indicates partial but not yet complete maturity across the assessed dimensions. In practical terms, this means that the pilot already benefits from a workable foundation for trustworthy AI operation, but that this foundation is not yet robust enough to absorb all relevant pressures without additional organisational strengthening.

The initial assessment is therefore shaped less by a widespread concentration of severe model-level risks and more by the quality of the surrounding operational context. The trustSense profile shows that the team performs well in collaboration, integrity, and technical proficiency, but remains weaker in areas that are especially relevant for stable risk control, such as privacy and compliance, resilience, and the ability to institutionalise improvements in mitigation practices. These weaker dimensions increase the likelihood that otherwise manageable issues may become harder to control over time.

This is particularly relevant for the current phase of LSP7 because the extracted trustGuard results show that the assessment has already pointed to response measures in several areas, while many of them were still marked as No action taken. As a result, the initial risk profile should not be read only in terms of inherent exposure, but also in terms of implementation maturity: part of the residual risk comes from the fact that recognition of needed controls is ahead of their practical deployment.

A key implication of this first assessment cycle is that the present LSP7 risk picture is mainly influenced by operational dependence on people, procedures, and governance routines. The initial level of risk is therefore not driven by evidence of widespread severe technical instability, but by the possibility that limited procedural maturity, incomplete documentation practices, and still-developing compliance routines may reduce the pilot's capacity to prevent, absorb, or recover from adverse situations.

For this reason, the initial risk assessment in LSP7 can be characterised as manageable, but requiring structured follow-up. The pilot does not appear to face an immediate generalised high-risk condition across the assessed trustworthiness properties, yet the current baseline remains sensitive to gaps in privacy readiness, resilience planning, and operational formalisation. This makes the present stage suitable for targeted corrective action before later phases of the pilot increase the need for stronger and more systematic trustworthiness controls.

## 3.6 Countermeasures

Based on the initial assessment results, the countermeasure strategy for **LSP7** was defined as a focused set of follow-up actions addressing the main weaknesses revealed by trustSense and the recommended controls proposed by trustGuard. At this stage, the intention is not to introduce a fully mature control framework in one step, but to prioritise a manageable first wave of measures that can reduce the most relevant residual risks in the current pilot setting.

A first group of countermeasures concerns **documentation and operational formalisation**. Since the trustGuard results associated the highest risk with **human error**, and explicitly proposed documentation-related actions for this area, a priority for LSP7 is to strengthen the way procedures, responsibilities, and expected model-use practices are recorded and communicated. In our context, this means moving from partially implicit working practices toward clearer written guidance for model operation, review, and incident handling.

A second priority concerns the **integration of the AI asset into more structured management processes**. trustGuard recommends including ML applications in asset-management processes and improving alignment with security policies, while the current extracted status shows that these actions were still marked as **No action taken**. For LSP7, this indicates the need to treat the clustering model less as a standalone technical component and more as an operational asset that requires traceable oversight throughout its lifecycle.

Another important area for countermeasures is **privacy and compliance reinforcement**. trustSense identified **Privacy and Compliance** as the weakest team dimension, with a score of **2.00**, and recommended foundational training on personal data handling together with simplified privacy-by-design guidance. In practical terms, this makes privacy-oriented awareness raising one of the most immediate and proportionate actions for LSP7, especially because it addresses a weakness already visible at team level before it develops into a broader operational problem.

The assessment also points to the need for stronger **resilience and recovery preparation**. trustSense assigned a score of **2.50** to **Resilience** and recommended the inclusion of resilience checkpoints in project planning together with clearer roles for recovery testing and incident documentation. For the pilot, this suggests that countermeasures should not focus only on

preventing faults, but also on improving the team's ability to respond in a structured way if something goes wrong during operation.

A further set of measures should target **threat awareness and mitigation discipline**. The trustSense results indicate only moderate maturity in threat awareness and limited evidence of established routines for introducing or updating mitigation actions, while trustGuard proposes several control measures that have already been identified but not yet activated. This combination suggests that LSP7 would benefit from lightweight but regular internal review points where newly identified risks, external developments, and pending controls are examined in a more systematic way.

For the technical asset itself, the proposed countermeasures should remain proportionate to the current risk picture. Since most trustGuard risk scores for the clustering model were low, the objective is not to redesign the asset or introduce heavy technical controls immediately, but rather to reinforce basic assurance mechanisms such as monitoring indicators for correct model functioning, periodic review of component vulnerabilities, and more explicit security-policy alignment. These actions would improve operational robustness without imposing unnecessary complexity on the current pilot phase.

Some countermeasures are also relevant for **fairness, interpretability, and privacy-related follow-up**, even if these areas did not emerge as the highest-risk domains in the initial assessment. trustGuard recommends actions such as regular bias regression testing, automated fairness testing within the development lifecycle, human-centred evaluation with operators, differential privacy-related measures, and awareness activities tailored to ML-specific risks, but the extracted report shows that these were still not activated. In LSP7, these measures can therefore be considered part of a second layer of controls to be introduced progressively after the most immediate documentation, compliance, and governance gaps are addressed.

From the LSP7 perspective, the countermeasure plan should therefore be phased. The first phase should prioritise low-complexity, high-relevance actions such as clearer documentation, privacy and compliance training, resilience planning, and the inclusion of the model in formal asset-management routines. A second phase can then address broader monitoring, fairness-related checks, and more advanced ML-specific controls once the operational baseline has become more stable.

The recommended countermeasures for LSP7 do not indicate a need for a complete technical redesign. Instead, they suggest adopting a more disciplined approach to managing trustworthiness around the deployed asset. At this stage, the key requirement is to translate the already identified recommendations into concrete pilot practices. This will help reduce the current risk profile through better documentation, heightened privacy awareness, improved preparedness for resilience, and clearer governance throughout the lifecycle of the clustering model.

# 4 AI System Evaluation

## 4.1 Quantitative metrics on pilot data

The quantitative evaluation of the Active Ageing pilot was conducted across three complementary dimensions during the initial pilot period. The first dimension covers the sociodemographic characterisation of the participant population, based on the baseline survey administered at enrolment and covering both elderly users and their relatives or informal caregivers. The second dimension addresses AI system performance, through a qualitative analysis of pilot data for a sample of 28 users, focusing on data coverage, the distribution of daily activity profiles across behavioural clusters, and the temporal transitions between clusters at the individual user level. The third dimension captures trust and explainability perceptions at an intermediate measurement point, based on a set of participatory Wooclap sessions conducted with a subsample of participants during the first semester of 2026. The remaining evaluation instruments — including the System Usability Scale[6] (SUS), the final XAI trust questionnaire, and the quality-of-life and impact assessments — are planned for the final measurement point and will be reported in D10.3. The metrics and analyses presented in this section, therefore, reflect the state of the evaluation at the end of the initial pilot phase.

### 4.1.1 Sociodemographic characterisation of participants

The baseline survey collected sociodemographic data from 42 elderly users, of whom 40 provided complete age and gender data. The age distribution of the elderly group reflects a notably advanced age profile, with a mean age of 85.9 years (SD = 5.0), a median of 85, and a range from 75 to 98 years. The majority of participants were in the 80–89 age band, accounting for 60% of the sample. In terms of gender, the elderly sample was predominantly female, with 36 participants (90%), and 4 male participants (10%).

Technology literacy was low: 27 participants (67.5%) reported no technology experience, 11 (27.5%) reported low familiarity, and only 2 (5.0%) reported medium familiarity. None reported high or expert-level technology use.

The frailty level of elderly participants was assessed using the Clinical Frailty Scale adapted for the FAITH pilot, and complete frailty data were available for 31 of the 42 registered elderly users. The distribution showed that 12 participants (38.7%) were classified as mildly frail (level 4), 11 (35.5%) as managing well (level 3), 7 (22.6%) as Agile (level 2), and 1 (3.2%) as severely frail (level 5). No participant scored at level 1 (Very agile) in the dataset. This profile confirms that the pilot is reaching its intended target population: older adults with reduced functional reserve who stand to benefit most from proactive monitoring and early anomaly detection, but who are not yet in a state of severe frailty that would preclude active participation.

The baseline survey also collected data from 86 relatives and informal caregivers linked to the pilot households, of whom 52 provided complete demographic records. The caregiver group presented a markedly different profile from the elderly users. The mean age of relatives was 62.1

[6] https://sus.tools/

years (SD = 6.5), with a range from 47 to 74, indicating that many caregivers are themselves in the early or mid-stage of older adulthood. The gender distribution among relatives was nearly equal, with 28 females (53.8%) and 24 males (46.2%) among respondents.

Technology familiarity was considerably higher among relatives, with 29 (55.8%) reporting medium familiarity, 16 (30.8%) high familiarity, and only 7 (13.5%) low familiarity. This profile is broadly consistent with expectations for a caregiver cohort in this age range and confirms that the caregiver-facing application interface is being used by a digitally capable population, which has positive implications for the quality of engagement data collected through the AI System.

### 4.1.2 AI System performance

The analysed dataset contains **3456 daily records** from **28 users**, covering the period from **2025-01-02** to **2026-02-08**. On average, each user contributes **123.4** labelled daily profiles, with a median of **97.5** days per user, indicating heterogeneous monitoring duration across participants.

The clustering output assigns each daily sleep-related mobility profile to one of four interpretable behavioural clusters. In this use case, we have identified pairs in clusters, where the first item refers to the transition speed from the bedroom to other rooms around bedtime, and the second to the transition speed on getting up. Accordingly, the labels are interpreted as follows: **Hummingbird = Fast-Fast**, **Sargeant = Fast-Slow**, **Earlybird = Slow-Fast**, and **Owl = Slow-Slow**.

*Table 6. Cluster identification*

| Transition pattern | Cluster Name | Description |
|---|---|---|
| Fast-Fast | Hummingbird | Energetic individuals active both during bedtime and waking routines, frequently moving in and out of their bedrooms, maintaining consistent activity and rhythm at both night and morning |
| Fast-Slow | Sargeant | Exceptionally disciplined sleepers exhibiting consistent and calm bedtime and waking habits, minimal bedroom noise, precise daily patterns similar to military-style routines, reflecting strong sleep regularity. |
| Slow-Fast | Earlybird | Individuals with quiet, stable bedtime habits but energetic morning routines, showing notable bedroom activity mainly upon waking, reflecting a proactive, organized start to their day. |
| Slow-Slow | Owl | People with quiet morning routines but active evenings, frequently leaving and returning to the bedroom at night, demonstrating dynamic pre-sleep behaviors such as showering or organizing tasks. |

The cluster descriptions used in LSP7 are the following using a bird analogy: **Hummingbird** describes energetic individuals active both during bedtime and waking routines; **Sargeant** corresponds to exceptionally disciplined sleepers with calm and regular bedtime and waking habits; **Earlybird** characterises users with quiet bedtime behaviour but energetic morning routines; and **Owl** refers to users with active evenings but quiet morning routines.

Across all labelled daily profiles, the cluster distribution is as follows: Hummingbird (Fast-Fast): 2543 daily profiles (73.6%); Earlybird (Slow-Fast): 673 daily profiles (19.5%); Sargeant (Fast-Slow): 140 daily profiles (4.1%); Owl (Slow-Slow): 100 daily profiles (2.9%). This distribution indicates which behavioural patterns are more prevalent in the monitored population and provides an aggregate view of the diversity of sleep-related indoor mobility captured by the pilot dataset.

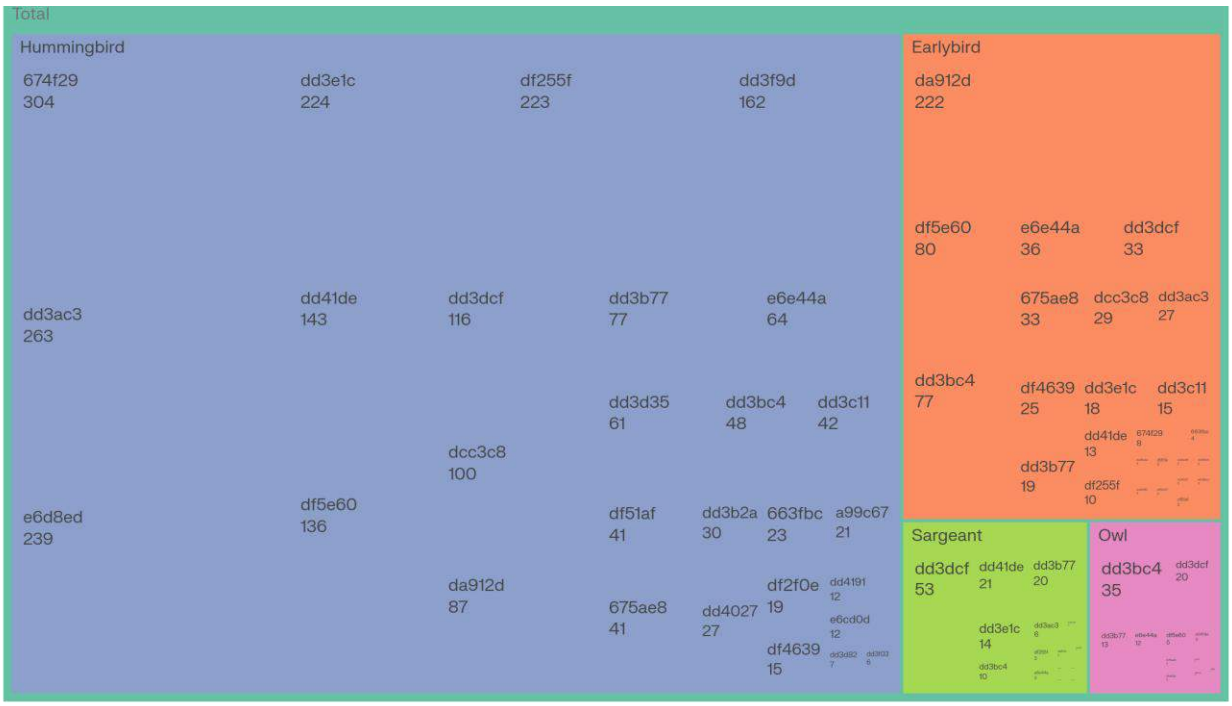



*Figure 8 Distribution users in clusters*

Beyond static distribution, the pilot data also support longitudinal analysis of behavioural changes. Considering consecutive observations for the same user, the dataset includes **3428 consecutive user-level transitions**, of which **946** correspond to a change from one cluster to another, yielding a transition rate of **27.6%**. The most frequent inter-cluster changes are: Earlybird to Hummingbird (302); Hummingbird to Earlybird (292); Sargeant to Hummingbird (67); Hummingbird to Sargeant (67); Hummingbird to Owl (47). These transitions are useful because they reveal whether behavioural patterns remain stable over time or fluctuate between distinct mobility phenotypes.

At participant level, **28 of 28 users (100.0%)** appear in more than one cluster during the observation period, showing that the method captures both stable and evolving sleep-related routines rather than assigning a single fixed label to each participant.

As an example of individual longitudinal analysis, user **67486880b3b5cefb77675ae8** was selected because this participant presents multiple observations and appears across all four behavioural clusters. This user contributes **80 daily profiles** spanning **4 clusters**, which makes the case suitable for illustrating temporal variability in behavioural patterns. The observed sequence can be used to identify stability, drift, and transitions between cluster states over time.

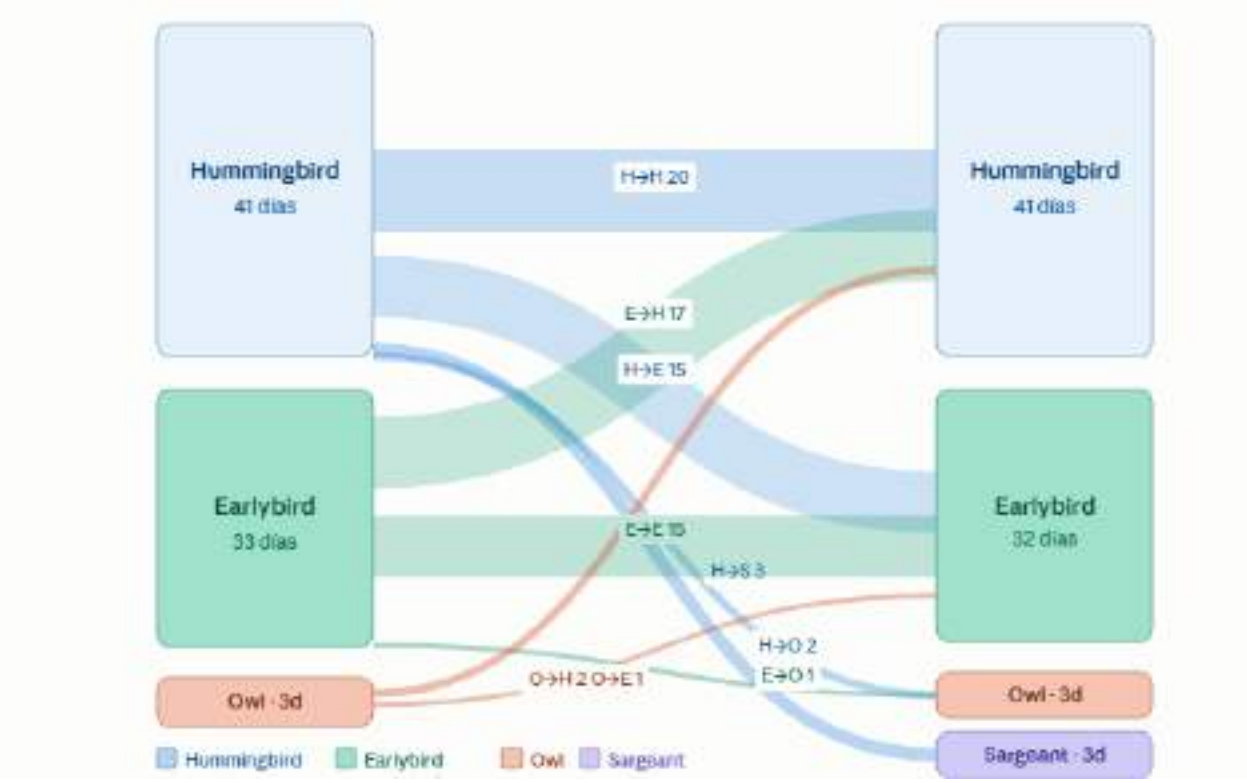



*Figure 9 Transition between clusters*

The diagram shows all day-to-day transitions for this user over the period (January 2025 to February 2026). The most frequent transitions are H→H (20 times) and H→E / E→H (15–17 times), which indicates that the user constantly alternates between Hummingbird and Earlybird.

Transitions toward Sargeant (3 times) and Owl (3 times) are rare and always originate from Hummingbird or Earlybird, never between each other. Owl and Sargeant are **transit** clusters: the user never remains in them for more than one day, with no self-transitions O→O or S→S.

The following graph shows the transitions over time:

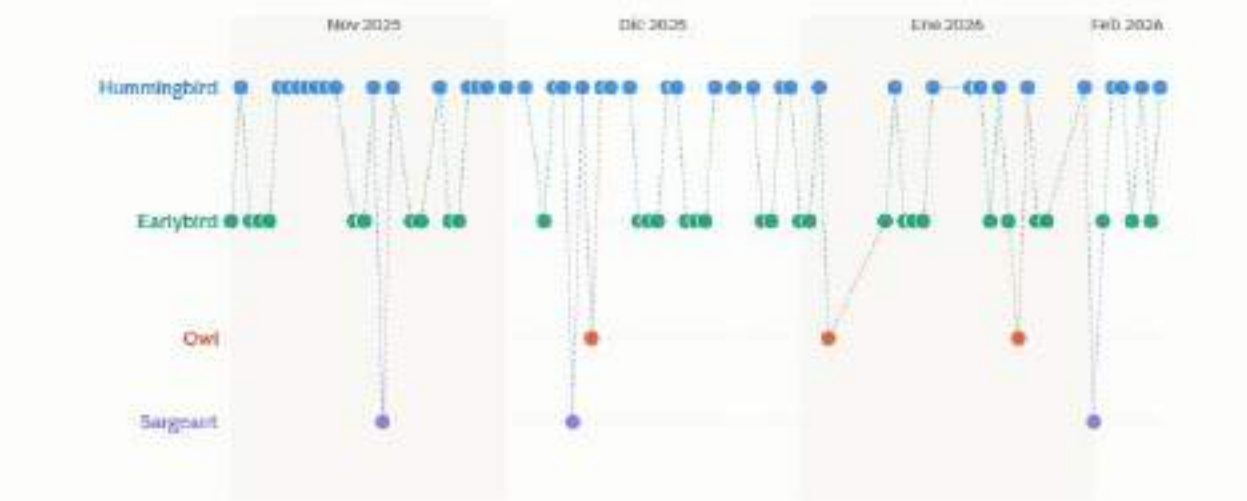



*Figure 10 Cluster classification per date*

Representative daily plots are provided to illustrate the four cluster archetypes produced by the algorithm. The image **Slow-Slow Owl** shows a profile with slow transitions both at bedtime and on rising. The image **Slow-Fast Earlybird** illustrates a pattern with slower evening transitions and faster morning behaviour. The image **Fast-Slow Sargeant** shows faster evening transitions and slower morning transitions. Finally, **Fast-Fast Hummingbird** represents users with rapid transitions in both periods. Together, these examples demonstrate how the fitted daily activity curves support an interpretable mapping from raw in-home mobility traces to clinically and behaviourally meaningful cluster labels.

**Fitted Function Slow-Slow OWL**

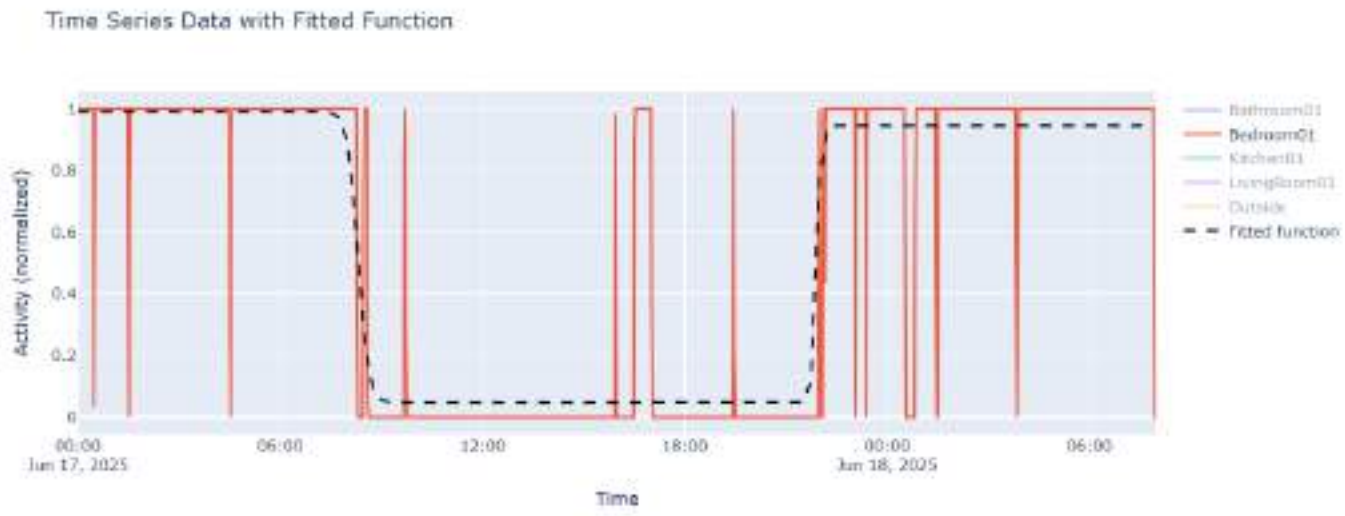



*Figure 11 Fitted Function Slow-Slow OWL*

## Fitted Function Slow-Fast Earlybird

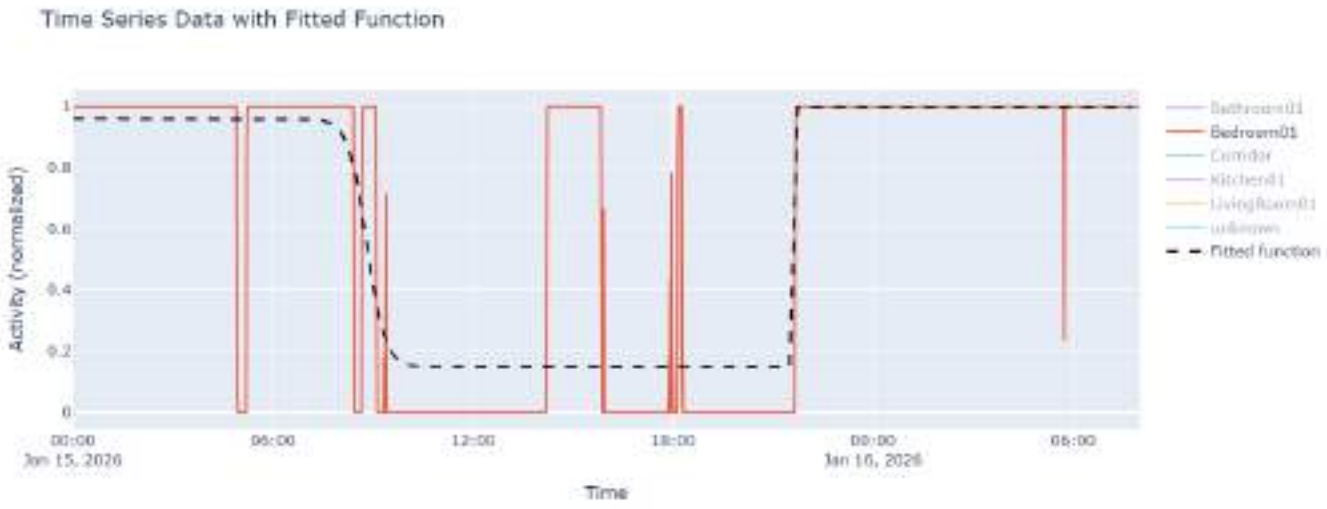



*Figure 12 Fitted Function Slow-Fast Earlybird*

## Fitted Function Fast-Fast Hummingbird

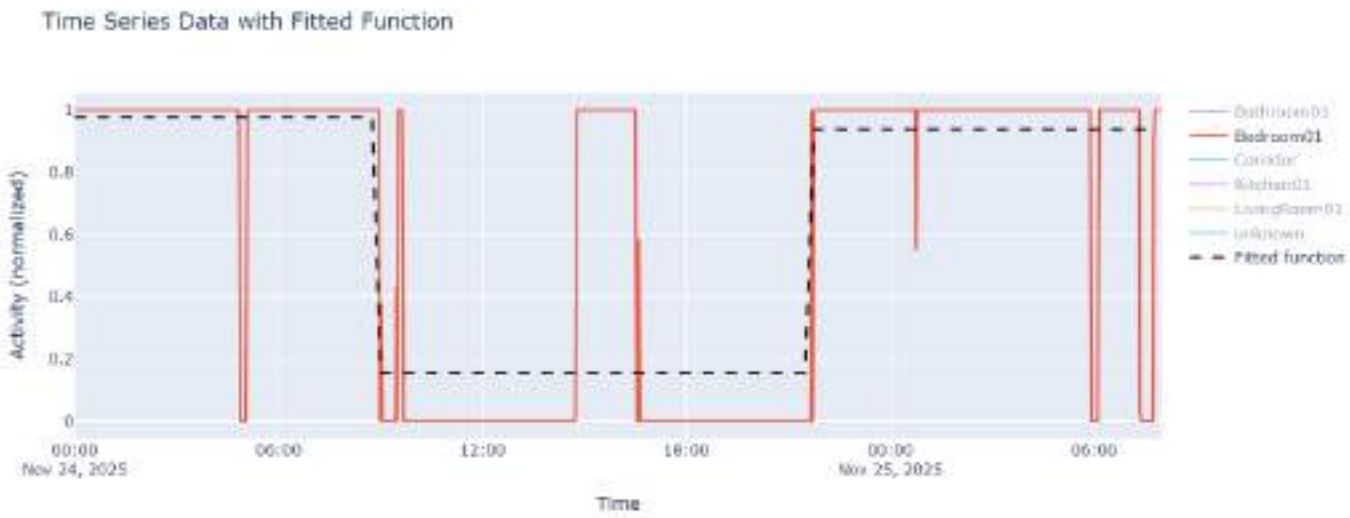



*Figure 13 Fitted Function Fast-Fast Hummingbird*

**Fitted Function Fast-Slow Sargeant**




*Figure 14 Fitted Function Fast-Slow Sargeant*

The quantitative results suggest that the LSP7 pilot data are appropriate for both population-level and individual behavioural analyses. The combination of aggregate cluster distribution, individual transition trajectories, and representative daily patterns offers a clear foundation for evaluating the clustering algorithm in a real active-ageing context. This information will also help inform later assessments of trustworthiness and impact in the pilot project.

### 4.1.3 Intermediate measurement of AI trust and perceptions

The intermediate evaluation was conducted through a series of participatory sessions on the Wooclap digital platform, attended by 39 valid respondents from the caregiver cohort across multiple sessions. In terms of age, 23 respondents (62.2%) were aged 66–75, 13 (35.1%) were aged 51–65, and 1 respondent was aged 76–85.

Technology familiarity followed a similar pattern than in the baseline: 51.4% reported medium familiarity, 29.7% high, 13.5% expert, and only 5.4% low familiarity. All respondents declared at least a general awareness of the concept of Artificial Intelligence, with 59.5% reporting a general idea and 37.8% reporting understanding of specific concepts.

Regarding prior use of technology for elderly care, 54.1% reported never using it, 29.7% used it regularly, and 13.5% used it occasionally. This confirms that a substantial portion of the session participants were engaging with AI-assisted monitoring concepts for the first time, making their trust perceptions particularly informative as a baseline for AI system acceptability.

Privacy was identified as a strongly salient value across the participant group: 75.7% considered data privacy to be "extremely important", and a further 24.3% considered it "somewhat important". No participant indicated that privacy was unimportant. This near-unanimous emphasis on privacy underscores the importance of the privacy-preserving design approach

adopted in the FAITH Active Ageing pilot and is consistent with findings from other LSPs in the consortium.

Participants were asked to rate their level of concern across six worry dimensions on a 0–10 scale. The results reveal that concerns about data misuse had the highest mean score (M = 8.58), followed closely by feelings of being surveyed (M = 8.08) and fear of system errors (M = 7.67). Concerns about costs and practical complications (M = 5.83) and general distrust of technology corporations (M = 5.75) were at a moderate level, while fear of being unable to uninstall the system attracted the lowest concern score (M = 3.75). The concentration of high scores on privacy- and accuracy-related dimensions is particularly relevant for the FAITH pilot, as it highlights the specific trust barriers that the AI features of the system must address to sustain participant engagement during the final pilot phase.

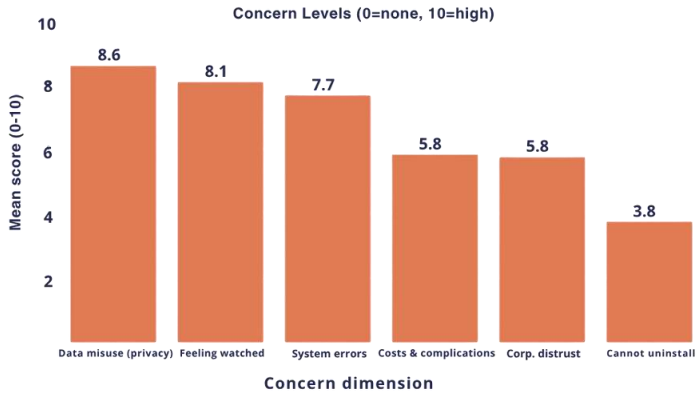



*Figure 15 Concern levels*

Finally, we collected perceptions of the AI system across the trustworthiness dimensions defined in FAITH AI_TAF . All dimensions received a mean score above the neutral midpoint of 3.0, indicating an overall positive disposition towards the system at the intermediate measurement point. Transparency obtained the highest mean score (M = 4.58), followed by Privacy (M = 4.25) and Robustness (M = 4.17). Equity and General Trust / Willingness to participate were rated equally (M = 3.83), while Explainability received the lowest mean score (M = 3.33). The relatively lower score for Explainability is consistent with the findings from the concern and trust-information items, and suggests that users have not yet fully internalised how the system's AI decisions are made and communicated. This gap between transparency (which relates to data governance and visibility of data flows) and explainability (which relates to the intelligibility of AI reasoning) points to a specific area for improvement in the interaction design of the caregiver-facing application before the final measurement point.

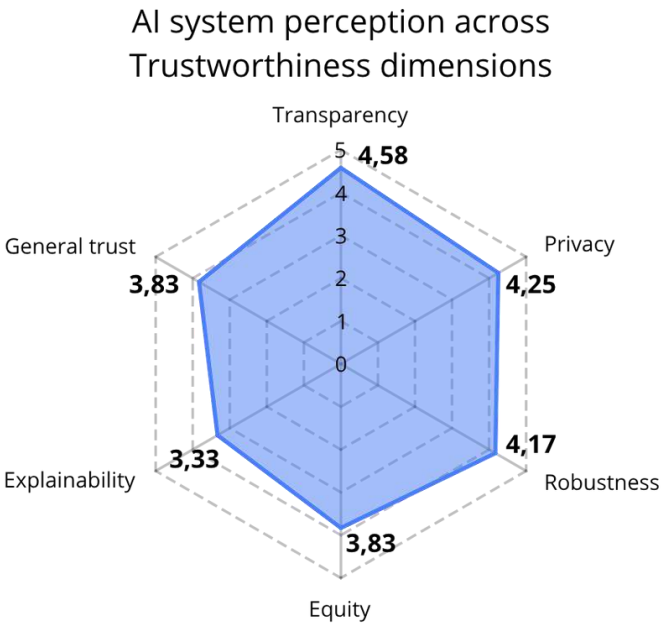



*Figure 16 AI system perception*

## 4.2 Trustworthiness Assessment

In LSP7, the trustworthiness assessment was carried out by combining the evidence produced by **trustGuard** and **trustSense**, in line with the FAITH methodological approach defined in D2.1, where trustworthiness is examined across a set of core dimensions covering both technical and socio-organisational assets. In practice, this allowed the pilot team to move beyond a purely model-centric reading and to assess how the deployed AI model and the technical users surrounding it are jointly performing against the relevant Trustworthy AI attributes.

For the LSP7 use case, not all seven dimensions carry the same weight. The assessment results indicate that the most informative and actionable dimensions are **Validity & Reliability**, **Security & Robustness**, **Transparency & Accountability**, **Explainability & Interpretability**, **Fairness**, and **Privacy**, because these are the areas where the available trustGuard and trustSense outputs provide direct evidence on model exposure, governance maturity, human performance, and operational readiness. The analysis below is therefore centred on those dimensions, while still retaining a brief comment on **Safety** to preserve consistency with the FAITH trustworthiness structure.

### 4.2.1 Validation & Reliability

According to the FAITH methodology, validity and reliability concern whether the AI system performs consistently and remains fit for its intended purpose under operational conditions. In LSP7, trustGuard provides the main evidence for this dimension by assessing the **LSP7 clustering model** in **production** and assigning it **medium criticality**, which already frames reliability as an operational rather than purely experimental concern.

The resulting picture is mixed but broadly acceptable at this stage. On the positive side, the extracted trustGuard results classify **failure or malfunction of the ML application** as **Low** risk, which suggests that the current model does not appear to exhibit a dominant immediate malfunction profile. At the same time, the same assessment shows that controls such as monitoring indicators for proper functioning, ongoing evaluation of components, and lifecycle-oriented oversight were identified but still marked as **No action taken**, which means that reliability is currently supported more by limited exposure and pilot conditions than by a mature continuous-assurance process.

trustSense complements this by showing that the AI technical team has strong **Technical Proficiency** (**4.00**) and good understanding of the intended use and expected outcomes of the AI system, with one questionnaire item receiving **5 - Strongly Agree**. This supports the conclusion that, in LSP7, validity and reliability are presently sustained by human competence and operational familiarity, but would benefit from more systematic monitoring and structured post-deployment verification mechanisms.

### 4.2.2 Security & Robustness

Security and robustness are central dimensions in the FAITH framework because they address the capacity of the AI system to withstand technical failures, hostile conditions, and other adverse events across the lifecycle. For LSP7, trustGuard gives a relatively contained picture in several classic technical threat areas, since **compromise of ML application components**, **poisoning**, and **oracle** are all assessed as **Low** risk.

However, the assessment does not support the conclusion that robustness is already mature. The vulnerability-related questions in trustGuard reveal that several expected safeguards, such as continuous assessment of the security level of ML components, alignment with organisational security-by-design processes, drift control services, cyber-resilience procedures, and formal documentation standards, were still answered negatively or left unresolved in the extracted report. In addition, many corresponding mitigation controls were proposed but remained at **No action taken**, indicating that the current robustness profile relies more on the modest present exposure of the model than on an already embedded security architecture.

trustSense reinforces this reading from the human side. The team shows good technical capability and collaboration, but only **moderate** scores in **Proactivity and Threat Awareness** (**3.33**) and **Resilience** (**2.50**), while incident recovery and prevention-of-recurrence behaviours were rated only **Neutral** in the questionnaire. Taken together, the tools suggest that LSP7 currently has a workable baseline for security and robustness, but one that remains vulnerable to procedural gaps and insufficiently formalised resilience practices.

### 4.2.3 Transparency & Accountability

Within D2.1, transparency and accountability are associated with traceability, documentation, explainable governance, and the ability to assign and demonstrate responsibility for AI-related actions and outcomes. In LSP7, the trustGuard outputs show that this dimension is less constrained by direct model opacity than by the surrounding governance context.

The strongest concern here is that the highest trustGuard risk for the LSP7 clustering model is **Human error**, rated **High**, with recommended controls focused on **documentation requirements for AI projects** and on including ML applications in **asset management processes**. This is highly relevant for accountability, because it indicates that the most significant present weakness is not a purely algorithmic one, but the risk of insufficiently structured human oversight, incomplete procedural recording, and unclear lifecycle ownership.

trustSense points in the same direction. The team received only **Moderate** results in **Responsibility and Ethics** (**3.67**), **Policy Adherence** (**3.00**), and several operational maturity dimensions, while one of the explicit recommendations is to create clearer responsibility matrices aligned with AI roles. For LSP7, this means that transparency exists only partially at the current stage: the pilot team understands the system and works collaboratively, but accountability would become stronger if documentation, role clarity, and formal governance arrangements were further consolidated.

### 4.2.4 Explainability & Interpretability

Explainability and interpretability in the FAITH approach refer to the extent to which AI outputs and their rationale can be understood by the relevant stakeholders, and whether this understanding is sufficient for meaningful human oversight. In the LSP7 evidence base, trustGuard does not identify explainability-related issues as dominant technical risks, since **reduced operator accountability due to lack of interpretability** and **reduced ability of operators to identify prediction errors** are both assessed as **Low** risk.

This suggests that, for the current pilot context, the model outputs are not perceived as inherently unusable or incomprehensible. Nevertheless, the recommended controls for these low-risk areas still include **human-centred evaluation with operators**, and these controls were marked as **No action taken**, which means that interpretability has not yet been systematically validated through structured user-facing assessment practices. In other words, low risk here should not be read as full maturity, but rather as the absence of a currently acute interpretability problem.

From the human asset side, trustSense provides indirect support for this dimension through the team's high **Technical Proficiency** (**4.00**) and strong awareness of how the AI system is intended to operate. For LSP7, the combined reading is that interpretability is presently adequate for operational use, but still requires stronger operator-centred validation and documentation if it is to mature from an acceptable condition into a demonstrably well-governed trustworthiness strength.

### 4.2.5 Safety

In the FAITH framework, safety concerns the potential harm that may arise when the AI system behaves incorrectly or is used in a way that negatively affects people, assets, or the operational environment. For LSP7, the available trustGuard results do not isolate safety as a dominant high-risk area in the way they do for **human error**, and the extracted assessment does not point to a broad cluster of high safety-related threat outcomes at model level.

Even so, safety remains indirectly relevant because the model is already assessed in **production** and because the overall criticality is **medium**. This means that safe operation in LSP7 currently

depends less on the existence of a single severe technical hazard and more on the ability of the surrounding technical team to use, supervise, and correct the system consistently.

The trustSense results show that the team has a solid technical base but only moderate resilience and only partial maturity in several governance-related traits. As a result, the LSP7 participants interpret safety as presently acceptable but closely linked to the strengthening of documentation, recovery planning, and organisational discipline, rather than as a dimension already guaranteed by the technical profile of the model alone.

### 4.2.6 Fairness

Fairness is defined in D2.1 as one of the core trustworthiness dimensions, covering the extent to which data, model behaviour, and decisions avoid unjustified bias and remain acceptable across relevant contexts and groups. In the LSP7 trustGuard assessment, **insufficient safeguards against bias** is rated **Low** risk, which indicates that fairness does not currently emerge as one of the dominant trustworthiness concerns for the assessed model.

However, the low risk score should be interpreted cautiously. trustGuard still recommends actions such as **regular bias regression testing**, **automated fairness testing within the model development lifecycle**, and the embedding of fairness criteria into **CI/CD pipelines**, yet all these measures are shown as **No action taken** in the extracted report. This means that fairness risk is currently judged to be limited, but also that the evidence base supporting that judgment is still relatively light and not yet backed by systematic ongoing checks.

trustSense adds a useful human-layer perspective. The technical team reports good data-quality handling capability and advanced technical competence, but only moderate maturity in problem solving, policy adherence, and structured improvement practices. For LSP7, this leads to the conclusion that fairness is not an urgent red-flag area in the first pilot cycle, yet it remains a dimension where stronger process discipline would improve confidence in future assessments.

### 4.2.7 Privacy

Privacy is one of the clearest dimensions of concern in the current LSP7 trustworthiness profile. While the trustGuard assessment rates **data disclosure** as **Low** risk and proposes privacy-related measures such as ML-specific awareness, differential privacy, and federated learning, all of these controls appear as **No action taken** in the extracted report. This indicates that the model-level exposure is not presently considered severe, but also that privacy protection is not yet supported by a strong active control layer.

The more significant signal comes from trustSense. The AI technical team scored only **2.00** in **Privacy and Compliance**, the lowest category in the assessment, and the associated recommendation explicitly calls for foundational training on personal data handling and simplified privacy-by-design guidance. This is directly reinforced by the questionnaire, where the statement on prioritising privacy and legal compliance received **2 - Disagree**, showing that privacy readiness is presently the least mature trustworthiness-related team capability captured in the LSP7 evidence.

## 4.3 Preliminary Impact Evaluation

From our LSP7 perspective, the AI system is already showing some clear impact, even though the pilot is still in an early stage. The system has the potential to support older people in staying at home for longer and to help caregivers notice changes in daily routines earlier. At the same time, we also see some limits and risks, especially around trust, usability, privacy, and the way the system is used in practice.

One positive impact is linked to **ageing in place**. The LSP7 solution uses sensors instead of video or audio, so it can monitor daily activity in a less intrusive way. This is important in our case because the aim is to support independent living without making people feel watched.

Another expected benefit is **early detection of changes**. The system looks at daily activity patterns, behaviour changes, and functional ability over time. In our pilot, this means that caregivers may notice possible health or frailty-related problems sooner, instead of waiting until the situation becomes more visible in daily life.

The system can also have a positive effect on **caregivers**. It gives alerts, trends, and AI-based insights that can help them understand what is happening in the home more quickly. This may reduce manual checking and make it easier to focus on the people who need attention most.

There is also a clear advantage in terms of **privacy**. Compared with camera- or microphone-based systems, the LSP7 setup is less invasive. That makes it more acceptable for older users and families, and it fits better with the trust requirements of this use case.

Still, the impact is not only positive. One important issue is **user trust**. Some older adults and caregivers may hesitate to use the system because they worry about data use, privacy, or AI decisions that are not easy to understand. If users do not trust the system, its practical impact will be limited, no matter how good the technical design is.

A second possible negative effect is **misuse or over-reliance**. If caregivers trust the alerts too much, they may give too much weight to AI outputs. If they trust them too little, they may ignore useful warnings. In both cases, the impact of the system becomes weaker than expected.

We also see a **privacy trade-off**. Even if the system is non-intrusive, it still collects sensitive behavioural data from the home. This means the value of the system depends not only on what it detects, but also on how well consent, transparency, and data protection are managed.

Another point is **usability**. Some older users may find digital tools difficult to adopt, especially if they are not used to this kind of technology. So part of the impact depends on training, support, and how easy the system is to understand and use in everyday life.

Our initial assessment suggests that the impact of the LSP7 system is promising but dependent on certain conditions. It has the potential to support independence, assist caregivers, and enhance early awareness of issues, but this is only achievable if trust, usability, and privacy are effectively addressed. In this regard, the system already demonstrates value; however, its true impact will rely on how seamlessly it integrates into daily caregiving routines.

# 5 Regulatory and Ethical Assessment

The deployment and evaluation of the AI System Under Test (AI SUT) within the Active Ageing LSP require a careful, structured assessment of the applicable regulatory and ethical frameworks governing the development and use of Artificial Intelligence in sensitive domestic and care-related environments. The AI SUT operates as a privacy-preserving home monitoring solution that processes sensor data from elderly users' homes to detect changes in daily activity patterns, generate behavioural segmentation outputs, and trigger anomaly-based alerts for relatives and caregivers. This operational context raises relevant considerations regarding data protection, the classification of the system under EU AI legislation, and the broader ethical implications of deploying AI-enabled monitoring in the personal living spaces of a vulnerable population.

The regulatory and ethical assessment in LSP7 is organised around three complementary dimensions: (i) the applicable data protection and certification pathway, with particular attention to the obligations arising from the processing of personal data in domestic and care settings; (ii) the system's alignment with the requirements of the EU Artificial Intelligence Act (EU AI Act), especially those pertaining to minimal and limited-risk AI systems operating in proximity to vulnerable persons; and (iii) the ethical considerations specific to the active ageing domain, including the protection of user dignity, autonomy, and privacy in a technology-mediated care environment.

The objective of this section is not to represent a formal regulatory submission, but to assess the extent to which the AI SUT and its deployment within the LSP7 environment are aligned with the regulatory expectations that would apply to its future operational scaling and to identify any areas where further compliance action is required before the replication phase.

## 5.1 Data Protection Pathway and Certification

The AI SUT deployed in LSP7 does not fall within the regulatory category of Software as a Medical Device (SaMD) under Regulation (EU) 2017/745[7] (EU Medical Device Regulation, MDR). The system does not perform diagnostic, prognostic, therapeutic, or clinical decision-making functions, nor does it produce outputs intended to inform clinical management. Its outputs consist of behavioural pattern summaries, daily activity indicators, and anomaly alerts intended for informal caregivers and family members, not for clinical professionals operating in a regulated diagnostic workflow. In consequence, CE Marking under the MDR framework is not required for the current configuration of the AI SUT as deployed in the initial pilot.

The applicable data-protection compliance framework for LSP7 is instead grounded in the obligations arising from the General Data Protection Regulation[8] (GDPR, Regulation (EU) 2016/679). Given that the AI SUT processes personal data in the form of sensor-derived behavioural information collected continuously in the private homes of elderly individuals, and given that this processing involves individuals who may be considered vulnerable, the conditions for conducting a Data Protection Impact Assessment (DPIA) under Article 35 GDPR are met. The

[7] http://data.europa.eu/eli/reg/2017/745/2026-01-01
[8] http://data.europa.eu/eli/reg/2016/679/2016-05-04

combination of systematic monitoring of a private space, the use of AI-based processing to infer behavioural states and anomalies from personal data, and the involvement of a vulnerable adult population all constitute risk-elevating factors that trigger the DPIA obligation.

A DPIA has been conducted in accordance with Article 35 GDPR. The assessment followed a structured process that included mapping all data flows associated with the AI SUT — from sensor acquisition at the home level, through cloud transmission and storage on Amazon Web Services (AWS) using server-side Advanced Encryption Standard (AES) encryption, to access control through the Vitana Web Portal and caregiver mobile application. The assessment confirmed that: (i) no audio or video data are collected, with all sensing performed via passive infrared, door, and key sensors; (ii) no data are sold or shared with third parties; (iii) data are processed under explicit informed consent, revocable at any time; (iv) access to personal data is governed by Role-Based Access Control (RBAC) mechanisms; and (v) data minimisation and storage limitation principles are observed throughout the data lifecycle.

The DPIA identified fourteen risks to data subjects' rights and freedoms, evaluated for likelihood and severity, yielding an inherent risk level prior to mitigation. The table below summarises the full risk register, the mitigation measures applied, and the resulting residual risk levels.

| Risk ID | Risk Description | Inherent Risk | Mitigation Measures | Residual Risk |
|---|---|---|---|---|
| **R1** | Unauthorised access to sensor data (external attacker or insider gains access to behavioural data revealing intimate daily life patterns) | HIGH | TLS 1.3 encryption in transit (MQTT, REST API); AES-256 at rest (AWS RDS, MongoDB, S3); RBAC; incident response procedure per GDPR Art. 33/34 | LOW |
| **R2** | Re-identification of pseudonymised data (linking event data to identity through temporal patterns or combined datasets) | HIGH | Pseudonymisation at point of collection; identity-code mapping stored in locked paper file accessible only to the Principal Investigator; separate encrypted database for identifiable data; data minimisation (no audio, video, GPS) | LOW |
| **R3** | Discriminatory profiling / algorithmic bias (AI model produces systematically inaccurate results for specific demographic groups, leading to inequitable care) | MEDIUM | Periodic bias audits monitoring false positive/negative rates by age, gender, and frailty level (planned at T1, T2 evaluation points); stratified recruitment; diverse training data; CEI-UPV ethics committee oversight | LOW |
| **R4** | False positives causing alarm and distress (frequent erroneous alerts cause anxiety in elderly users or relatives and lead to unnecessary interventions) | HIGH | Configurable alert thresholds (conservative defaults); SHAP-based explainability layer accompanying every alert; caregiver training on alert interpretation; co-design workshops integrating user feedback into alert design | MEDIUM |
| **R5** | False negatives causing missed interventions (AI fails to detect a real health event, leading to delayed response and harm) | HIGH | Conservative detection defaults; mandatory human-in-the-loop validation protocol (caregivers treat AI alerts as decision support, not directives; continuous | MEDIUM |

| | | | | |
|---|---|---|---|---|
| | | | performance monitoring; defined escalation protocols) | |
| **R6** | Surveillance chilling effect (elderly users modify behaviour due to perceived monitoring, reducing well-being) | MEDIUM | Non-intrusive sensor technology; right to withdraw at any time without consequence; co-design workshops; transparent communication materials | LOW |
| **R7** | Inadequate or uninformed consent (participants do not fully understand what they are consenting to, due to cognitive or literacy barriers) | HIGH | Adapted consent procedure; verbal explanation; recruiters trained in communication with elderly; copy provided to participant; DPO review of consent materials | LOW |
| **R8** | Data breach or loss during cloud storage or transmission | HIGH | AES-256 at rest; TLS 1.3 in transit; AWS EU-region hosting; automated backups; documented disaster recovery procedures | LOW |
| **R9** | Scope creep / secondary use of data by consortium members, insurers, or third parties | MEDIUM | Strict Data Use Agreements (DUAs) for all external data sharing; RBAC preventing access beyond role-defined scope; only aggregated/anonymised results published externally; purpose limitation documented | LOW |
| **R10** | Dependency / caregiver over-reliance on AI alerts, reducing professional judgement | MEDIUM | SHAP explanations contextualising every alert; caregiver on-boarding training on AI limitations; human-in-the-loop validation protocol; alert acknowledgement rate monitoring planned (threshold: <70% triggers threshold review) | LOW |
| **R11** | Model drift / performance degradation (models do not generalise to new users or contexts over time) | MEDIUM | MLflow and DVC for model version control and dataset lineage; continuous monitoring of precision/recall, DBCV, and SHAP stability; periodic model retraining; foundational model architecture planned for replication phase | LOW |
| **R12** | Data subject rights obstruction (difficulty exercising access, rectification, erasure, or portability rights due to pseudonymisation challenges) | LOW | Pseudonymisation key managed by Principal Investigator; documented rights-request procedure; right-to-erasure tested end-to-end before pilot launch (per DPIA Recommendation REC-4) | LOW |
| **R13** | Cross-border data transfer risks (AWS infrastructure failure to remain within EU jurisdiction) | MEDIUM | AWS EU-region selection confirmed; Data Processing Agreement (DPA) with AWS Europe SARL covering all processing activities and sub-processor list | LOW |
| **R14** | Bias in training data (historical ACTIVAGE/Vitana datasets underrepresent certain user groups) | MEDIUM | Dataset curation and representativeness assessment; bias testing during model training; ACTIVAGE and Vitana data validation; stratified pilot recruitment | LOW |

Across all fourteen identified risks, twelve have been reduced to LOW residual risk following implementation of the technical and organisational measures described above. Two risks — R4

(false positives) and R5 (false negatives) — remain at a MEDIUM residual level due to inherent limitations in AI model performance in a real-world deployment and are managed through human oversight protocols and continuous monitoring rather than being fully technical in nature.

Pursuant to Article 36 GDPR, prior consultation with the supervisory authority is therefore not required prior to processing. The system's compliance with GDPR and its privacy-by-design architecture are further supported by the AI SUT's Data Passport registered in the FAITH AI Model Hub, which documents data provenance, collection methodology, and the privacy conditions established in the DPIA, ensuring end-to-end traceability and auditability of the data governance framework.

## 5.2 Alignment with the EU AI Act and Ethical Considerations

The AI SUT of LSP7 has been assessed in the context of the EU Artificial Intelligence Act (Regulation (EU) 2024/1689)[9], which entered into force on 1 August 2024 and applies progressively according to a phased implementation timetable, and establishes a risk-based regulatory framework for AI systems placed on the EU market. As part of the Cartography phase of the FAITH AI_TAF v1 methodology applied in Chapter 3, an initial classification of the system was conducted in accordance with the criteria set out in Article 6 and Annex III of the AI Act.

Based on the use case currently assessed, the AI SUT does not appear to fall within the high-risk AI system categories enumerated in Annex III of the AI Act. This is a preliminary assessment and may need to be revisited if the system's intended use or deployment context changes. It is not used for biometric identification, assessment of individuals in education or employment contexts, essential private or public services, law enforcement, migration, or administration of justice. It does not make autonomous decisions with legal or similarly significant effects on the individuals it monitors. The system's anomaly detection outputs will be exclusively advisory; they will be communicated to caregivers and relatives who retain full autonomy to assess, validate, or disregard alerts, and no automated action is taken on the basis of AI outputs alone. This human-in-the-loop design is a fundamental architectural feature of the system and supports meaningful human oversight, consistent with the oversight principles set out in Article 14 of the AI Act. The system is therefore classified as a minimal-risk under the current EU AI Act framework.

As a system that processes personal data and generates outputs potentially related to the health and well-being of elderly individuals, the AI SUT nonetheless operates in a sensitive domain and must observe heightened transparency obligations under Article 50 of the AI Act. In particular, users and their representatives must be clearly informed that they are interacting with an AI-enabled monitoring system, and the logic underlying anomaly detection outputs should be made comprehensible to non-expert users. The Vitana AI-Enhanced Application addresses these requirements through an explainability interface that contextualises anomaly alerts in terms of deviations from the individual user's established behavioural baseline, an approach that the intermediate trust measurement (Section 4.1) confirmed is preferred by end users over purely statistical or technical justifications.

[9] http://data.europa.eu/eli/reg/2024/1689/oj

**Risk management and quality assurance**. Although the system is not classified as high-risk under the current AI Act framework, the FAITH approach treats the application of a comprehensive risk management methodology as a fundamental requirement for responsible AI deployment, irrespective of formal risk class. The FAITH AI_TAF v1 has been applied systematically throughout the initial pilot to identify assets, analyse threats, assess vulnerabilities, estimate risks, and define countermeasures across the seven trustworthiness dimensions of the framework. Key threats identified in the active ageing context, including model drift due to behavioural evolution over time, false anomaly alerts leading to caregiver alarm fatigue, and privacy risks from unauthorised access to domestic sensor data, have been mitigated through a combination of technical controls and organisational procedures, as documented in Chapter 3. This structured risk governance approach is directly aligned with the quality assurance principles that the AI Act applies to high-risk systems and that FAITH, as a matter of best practice, extends to all LSP deployments.

**Data governance and dataset quality**. The AI Act imposes stringent requirements on data governance, particularly for high-risk systems, emphasising the need for data that is relevant, representative, and free of harmful biases. In the active ageing context, the SUT's anomaly detection models are personalised to each individual elderly user's behavioural profile, derived from that user's own historical sensor data. This personalisation-by-design approach significantly reduces population-level bias risk, as the system does not benchmark any individual against a generalisable demographic norm, but rather against their own established routine. Data quality is monitored through the Data Version Control (DVC) pipeline integrated into the AI development workflow, and Data Passports registered in the FAITH AI Model Hub ensure that data provenance and preprocessing steps are fully documented.

**Transparency and human oversight**. Transparency obligations are fulfilled through multiple mechanisms. The informed consent process conducted at enrolment provides participants with clear information on the nature, purpose, and scope of data processing. The Vitana Web Portal and caregiver application make AI outputs visible and interpretable to non-expert users, avoiding the opacity of a black-box system. The sessions conducted during the intermediate evaluation phase (described in Section 4.1) have confirmed that users place high value on professional-validated and baseline-comparison forms of explanation, and this feedback has directly informed the design of the explainability features of the Vitana AI-Enhanced Application currently in internal testing. The system's overall transparency posture received the highest mean trust score (M = 4.58) in the intermediate evaluation, confirming that users perceive the system's data governance practices as trustworthy.

**Ethical considerations for AI deployment in the active ageing domain**. The deployment of AI-enabled monitoring in the private homes of elderly individuals raises ethical considerations that extend beyond formal regulatory compliance and require explicit attention within the FAITH LSP7 framework. Three dimensions are particularly relevant.

- **User dignity and autonomy**. The right of older adults to control their own living environment and to make informed, freely given choices about technology-mediated monitoring is a central ethical commitment of the pilot. All participation is voluntary, all consents are fully informed and individually revocable, and the system is designed to

minimise intrusiveness by using only passive, non-audiovisual sensors. The ethical dimension of autonomy is also reflected in the co-creation sessions' findings: 54.1% of participants reported no prior experience with care technologies, underscoring the importance of ensuring that enrolment decisions are made with a genuine understanding of the system's capabilities and limitations rather than deferring to caregivers or service providers.

- **Privacy in the domestic sphere**. The home is a space with a particularly strong expectation of privacy in European legal and ethical traditions. The pilot addresses this through a privacy-by-design architecture that avoids collecting any data that could enable direct observation of intimate behaviour, excludes audio and video capture entirely, and retains sensor event data only for the period necessary for modelling and alert generation. The near-unanimous emphasis on data privacy among the session participants (75.7% rated it "extremely important") confirms that this ethical concern is actively salient for the user community and that continued attention to privacy communication and governance will be essential throughout the replication phase.
- **Algorithmic fairness and non-discrimination**. Given the demographic composition of the pilot's elderly user group — predominantly female (90%), with low educational attainment and very limited prior technology exposure — it is important to assess whether the system's AI outputs might disadvantage users whose behavioural profiles deviate from those that the models were trained on, or whose frailty levels affect the rate at which meaningful anomalies can be detected. The FAITH AI_TAF fairness assessment conducted via trustGuard has not identified structural bias risks in the current personalised anomaly detection architecture, as the model adapts to each user's individual baseline. However, as the system expands to a more diverse user population during the replication phase, including new deployment sites with potentially different cultural routines and household configurations, ongoing fairness monitoring and cross-cohort validation will be required to ensure equitable performance across the full range of participants.

# 6 Conclusions and roadmap for the final pilot phase

The initial pilot of the Active Ageing LSP has confirmed that a privacy-preserving, AI-enabled home monitoring solution can be successfully deployed and operated in real-life domestic environments involving older adults and their informal caregivers. The reporting period covered by this deliverable has advanced the pilot from its preparatory phase into full operational execution in Valencia, producing a structured body of evidence on deployment feasibility, trustworthiness, user perceptions, and regulatory alignment. This final chapter draws together the main achievements of this phase and sets out the forward plan and risk landscape for the replication phase.

## 6.1 Summary of achievements in the second period

The initial pilot has delivered on all principal objectives established in the Description of Action and in D10.1. The following achievements reflect the pilot's state at the time of reporting.

- **Operational deployment**. The sensor-based monitoring infrastructure has been successfully deployed across 55 homes in Valencia, with 175 users in active operation, 55 elderly individuals and 120 relatives or informal caregivers. The complete data collection pipeline, from passive sensor event acquisition through the Dagster ETL orchestration layer to cloud-based storage on AWS, is fully operational. Live data processing provides a continuous stream of sensor events suitable for AI model consumption, with approximately 9,000 events accumulated since the first installations in January 2025.
- **AI system integration**. The two core AI components of the System Under Test have been integrated in the technical infrastructure. The behaviour segmentation model (HDBSCAN/UMAP/DBCV, TRL 5) is integrated within the caregiver mobile application, with cluster predictions available to caregivers in daily use. The anomaly detection model (2-sigma/XGBoost, TRL 3) has advanced through iterative validation and refinement during the reporting period. The Vitana AI-Enhanced App, which delivers AI-based insights through a dedicated caregiver interface incorporating SHAP-based explainability, has been released for internal testing and is on track for broader release in the replication phase.
- **FAITH framework application**. The FAITH AI_TAF v1 methodology has been applied in full across the six phases of the Cartography-Threat-Impact-Vulnerability-Risk-Countermeasure cycle, producing a structured trustworthiness risk profile for LSP7. trustSense, trustGuard, and the AI Model Hub have been deployed and validated within the pilot context, with Data Passports registered for the declared AI assets (M1, M2, DA1, DA2). Formal tool feedback has been provided to the WP2 tool development teams.
- **User engagement and trust evidence**. End-user co-creation workshops were held at the Senior University of Polytechnic University of Valencia with 150 attendees. The intermediate measurement conducted through Wooclap sessions engaged 39 valid respondents and produced structured evidence on AI trust perceptions, privacy salience, concern profiles, and preferred explanation formats. Overall trustworthiness perceptions at the intermediate measurement point are positive across all dimensions assessed, with

transparency receiving the highest mean score (M = 4.58) and explainability identified as the primary area for continued design attention (M = 3.33).

- **Regulatory and ethical compliance**. Formal ethical approval has been granted by the Ethics Committee of the Universitat Politècnica de València (CEIUPV). The DPIA has been completed. Data governance arrangements, including RBAC, AES encryption, pseudonymisation, and explicit informed consent procedures, are fully operational. The system has been assessed as minimal-risk under the EU AI Act and confirmed as non-applicable to the MDR/SaMD framework, consistent with the medium-criticality characterisation from the Cartography phase.
- **KPI progress**. At the current stage of the pilot, the user engagement and adoption targets remain work in progress pending the full release of the Vitana AI-Enhanced App to the pilot user base. The clustering accuracy and anomaly detection performance KPIs will be formally assessed against their quantitative targets (DBCV > 0.8; precision/recall > 85%) in the final evaluation phase. The trustworthiness validation KPI (application of the FAITH AI_TAF , documented in this deliverable) is fulfilled. The user trust perception KPI (≥75% reporting high trust in transparency and reliability) is on track based on intermediate measurement results.

## 6.2 Plan for Replication Phase

The replication pilot of the Active Ageing LSP runs from M20 to M51 in accordance with the Description of Action, with replication user recruitment starting in May 2026 and data collection and validation activities extending through March 2028. The replication phase has four defining characteristics relative to the initial pilot: geographic expansion, AI system advancement, framework consolidation, and scale.

The replication phase extends the pilot beyond Valencia to additional municipalities and regions in Spain, with the municipality of Alaquàs identified as the primary candidate for the first replication site. Preparatory stakeholder consultations have been conducted with the Valencia Regional Government, municipal social services, and service provider networks. At least five service provider partners have been identified to support recruitment in replication areas. A session with a professional women's retiree association in Madrid confirmed interest in expanding the geographic reach beyond the Valencia region.

The replication phase will integrate a foundational model approach for anomaly detection, designed to address the well-recognised challenge of data scarcity at new sites during the early deployment period. Current personalised anomaly detection models require a sufficient volume of user-level behavioural data to generate reliable outputs — a condition not fully met at the outset of deployment in new locations. The adoption of foundational models is expected to accelerate model adaptation to new user cohorts and to support more consistent anomaly-detection performance across sites with differing environmental and behavioural profiles.

An additional 250 users (elderly individuals and their caregivers) will be recruited in the replication phase, following the same multi-channel recruitment framework proven effective in the initial pilot: digital campaigns, senior centre engagement, social services collaboration, and

service provider partnership networks. The participant profile remains consistent with the initial pilot: older adults aged 65 and over living alone or spending most of the day at home, and their relatives or informal caregivers.

The FAITH AI_TAF v2 methodology will be applied in the replication phase, enabling the transition from asset-level assessment conducted in the initial pilot to system-level assessment of the broader socio-technical environment, as planned from M33 onwards. The System Trust Modeller (STM) will also be integrated into the replication phase, operationalising the FAITH AI_TAF in a more dynamic, automated risk-monitoring mode.

## 6.3 Identified risks for the Replication Phase

Building on the lessons learned from the initial pilot and on the risk profile produced by the FAITH AI_TAF application in Chapter 3, the following risks have been identified as requiring active monitoring and management during the replication phase.

*Table 7. Risks for replication phase*

| # | Risk | Likelihood | Potential Impact | Mitigation |
|---|---|---|---|---|
| R1 | **Data scarcity for anomaly detection at new sites** | High | Medium — reduces model reliability in early deployment period | Adopt foundational model architecture for anomaly detection; plan pre-deployment data familiarisation period |
| R2 | **Recruitment shortfall in replication areas** | Medium | Medium — delays reaching 250-user target | Activate identified service provider partnerships early; extend digital campaign strategy proven effective in initial pilot |
| R3 | **Model drift due to behavioural heterogeneity across sites** | Medium | Medium — affects anomaly detection precision and caregiver trust | Implement continuous model monitoring via DVC pipeline; include cross-site validation in the evaluation plan |
| R4 | **Caregiver alarm fatigue from false anomaly alerts** | Medium | High — undermines adoption and trust in system outputs | Complete internal testing and validation of Vitana AI-Enhanced App before broad release; calibrate alert thresholds using pilot data |
| R5 | **Participant dropout due to technology concerns** | Low-Medium | Medium — reduces data continuity and evaluation quality | Maintain personalised onboarding and support; use trust-building communication strategy informed by AI session findings |
| R6 | **Organisational dependency on key personnel** | Low | Medium — risk to continuity if key AOA/BRIDG staff unavailable | Document processes; cross-train team members; maintain replication plan documentation at consortium level |

These risks do not fundamentally challenge the viability of the replication phase. The mitigations identified are either already in place from the initial pilot experience (R2, R6) or are directly addressed by planned technical decisions (R1, R4). The risk profile of LSP7 in the replication phase is characterised by medium operational complexity rather than by structural technical or regulatory barriers, and the groundwork established during the initial pilot — in terms of ethical clearance, data governance, AI infrastructure, user engagement methodology, and FAITH tool integration — provides a solid foundation for the transition to the replication phase.

# 7 References

The references in the current document are presented as footnotes.