



TRUSTEE

TRUSTEE D3.1. TRUSTEE Core & Fair Knowledge

Work package	WP3: TRUSTEE Core & Fair Knowledge
Task	Task 3.1 TRUSTEE Homomorphic enabled data fusion Task 3.2 AI model as a Service Task 3.3 Certification and auditing Task 3.4 TRUSTEE Knowledge Repository Task 3.5 EIDAS, Authentication service, Sensor Enabler Task 3.6 Explainable AI functions
Authors	Emmanouil Spanakis – FORTH, Vangelis Sakkalis – FORTH, Stelios Sfakiannakis – FORTH, Cristina Regueiro – TECNALIA, Marisa Escalante – TECNALIA, Julen Bernabe – TECNALIA, Iñaki Seco – TECNALIA, Santiago De Diego – TECNALIA, Ivan Gutierrez – TECNALIA, Eneko Gomez – TECNALIA, Evangelos Markakis – HMU, Dimitra Papatsaroucha – HMU, Maria Zotou – ACCELIGENCE, Stergios Zissakis – ACCELIGENCE, Stelios Vourakis – ACCELIGENCE, Ilias Politis – INQBIT, Athanasios Vasilios Grammatopoulos – INQBIT, Ioannis Stylianou – INQBIT, Alexandros Fotiadis – INQBIT, Michał Kępkowski – INQBIT, Konstantinos Kefalas – INQBIT, Maria Liatsou – INQBIT, Maria Sampazioti Vasiliki – INQBIT, Giulia Rutto – RINA-C, Manuele Barbieri – RINA-C, Katarzyna Barud – UNIVIE, Nikolaus Forgó – UNIVIE, Clara Saillant – UNIVIE, Hande Özkayagan-Prändl – UNIVIE, Sarwar Sayeed – ENU, Nikolaos Pitropakis – ENU, Panagiotis Radoglou Grammatikis – K3Y, Savvas Ouzounidis – K3Y, George Nakas – K3Y, Marios Siganos – K3Y, Evangelos Mourikis – K3Y, Zacharenia Lekka – K3Y, Antonios Sarigiannidis – K3Y
Dissemination level	PU
Status	Pre-final
Due date	30/11/2023
Document date	30/11/2023
Version number	1.0



Funded by the
European Union

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. Neither the European Union nor the granting authority can be held responsible for them.

TABLE OF CONTENTS

Table of contents	3
1. Executive Summary	17
2. Introduction	18
2.1. Purpose and scope.....	19
2.2. Relation with other Work Packages and Deliverables.....	19
2.3. Approach and structure	20
3. PART I: Homomorphic enabled data fusion (HEDF).....	23
3.1. INTRODUCTION	23
3.1.1. Need and Objectives	23
3.1.2. Innovation	23
3.2. BACKGROUND	23
3.2.1. Data Acquisition.....	23
3.2.2. Definition of data abstraction.....	24
3.2.3. Definition of a unified indexing framework	26
3.2.4. Importance of scalability, responsiveness, and reliability	27
3.2.5. Homomorphic Enabled Data Encryption	28
3.3. REQUIREMENTS.....	32
3.4. ARCHITECTURE	36
3.4.1. High level	36
3.4.1. Low level.....	37
3.5. IMPLEMENTATION	44
3.5.1. Functional Description.....	44
3.5.2. Operational Description	47
3.5.3. Technical Description	49
3.5.4. Technical Details.....	52
3.6. VALIDATION.....	53
3.6.1. Encryption Layer.....	53

TRUSTEE D3.1 TRUSTEE Core & Fair Knowledge

3.7.	NEXT STEPS	59
3.8.	CONCLUSIONS.....	59
4.	PART II: AI model as a Service AlmaaS	60
4.1.	INTRODUCTION	61
4.1.1.	Need and Objectives	61
4.1.2.	Innovation	63
4.2.	BACKGROUND	65
4.3.	REQUIREMENTS.....	67
4.4.	ARCHITECTURE	71
4.4.1.	High level	71
4.4.2.	Low level.....	71
4.5.	IMPLEMENTATION.....	74
4.5.1.	Functional Description.....	74
4.5.2.	Operational Description	75
4.5.3.	Technical Description	82
4.5.4.	Technical Details.....	90
4.6.	VALIDATION.....	92
4.7.	NEXT STEPS	103
4.8.	CONCLUSIONS.....	104
5.	PART III: Trustworthiness AI Support Design Framework (TAI-SDF)	105
5.1.	INTRODUCTION	105
5.1.1.	Need and Objectives	105
5.1.2.	Innovation	105
5.2.	BACKGROUND	106
5.2.1.	Trustworthiness-by-design frameworks for monitoring, auditing and conformity check	106
5.2.2.	Security Metrics For AI Models	107
5.3.	REQUIREMENTS.....	108
5.4.	ARCHITECTURE	111
5.4.1.	High level	111

TRUSTEE D3.1 TRUSTEE Core & Fair Knowledge

5.5.	IMPLEMENTATION	112
5.5.1.	Functional Description	112
5.5.2.	Operational description	113
5.5.3.	Technical Description	116
5.6.	VALIDATION.....	118
5.7.	NEXT STEPS	119
5.8.	CONCLUSIONS.....	119
6.	PART IV: Accountable Transaction Recorder (ATR).....	120
6.1.	INTRODUCTION	120
6.1.1.	Need and Objectives	120
6.1.2.	Innovation	120
6.2.	BACKGROUND	120
6.2.1.	Blockchain Technology	120
6.2.2.	Hyperledger Fabric.....	121
6.2.3.	Consensus Mechanism.....	121
6.2.4.	Containerisation	122
6.2.5.	Public Key Infrastructure	122
6.2.6.	State Database	122
6.2.7.	Policies	123
6.2.8.	Immutability	123
6.2.9.	Private Data Collection	124
6.3.	REQUIREMENTS.....	125
6.4.	ARCHITECTURE	128
6.4.1.	High level	128
6.4.2.	Low level.....	128
6.5.	IMPLEMENTATION.....	129
6.5.1.	Functional Description.....	129
6.5.2.	Operational Description	132
6.5.3.	Technical Description	133

TRUSTEE D3.1 TRUSTEE Core & Fair Knowledge

6.5.4.	Technical Details.....	133
6.6.	VALIDATION.....	134
6.6.1.	Channel Creation and network components	134
6.7.	NEXT STEPS	137
6.8.	CONCLUSIONS.....	137
7.	PART V: Knowledge Repository (KR)	138
7.1.	INTRODUCTION	138
7.1.1.	Need and Objectives	138
7.1.2.	Innovation	139
7.2.	BACKGROUND	139
7.2.1.	InterPlanetary File System (IPFS)	139
7.2.2.	Private IPFS Network	139
7.2.3.	Private IPFS Advantages.....	140
7.2.4.	Content-Addressable Storage (CAS)	140
7.2.5.	Distributed Hash Table (DHT).....	140
7.2.6.	Libp2p	141
7.2.7.	MerkleDAG (Directed Acyclic Graph).....	141
7.2.8.	Content Identifier (CID).....	141
7.2.9.	InterPlanetary Naming System (IPNS)	142
7.2.10.	Bitswap Protocol	142
7.3.	REQUIREMENTS.....	144
7.4.	ARCHITECTURE	146
7.4.1.	High level	146
7.4.2.	Low level.....	148
7.5.	IMPLEMENTATION	149
7.5.1.	Functional Description.....	150
7.5.2.	Operational Description	153
7.5.3.	Technical Description	154
7.5.4.	Technical Details.....	154

TRUSTEE D3.1 TRUSTEE Core & Fair Knowledge

7.6.	VALIDATION.....	156
7.7.	NEXT STEPS	159
7.8.	CONCLUSIONS.....	159
8.	PART VI: Authentication Manager (AM).....	159
8.1.	INTRODUCTION	160
8.1.1.	Need and Objectives	160
8.1.2.	Innovation	160
8.2.	BACKGROUND	161
8.2.1.	Self-Sovereign Identity (SSI).....	161
8.2.2.	SSI IN GAIA-X.....	161
8.3.	REQUIREMENTS.....	162
8.4.	ARCHITECTURE	164
8.4.1.	High level.....	164
8.4.2.	Low level.....	166
8.5.	IMPLEMENATATION.....	171
8.5.1.	Functional Description.....	171
8.5.1.	Operational Description	175
8.5.2.	Technical Description	181
8.5.3.	Technical Details.....	183
8.6.	VALIDATION.....	188
8.6.1.	Validation Methodology	188
8.6.2.	Validation & Unit Testing.....	188
8.7.	NEXT STEPS	189
8.8.	CONCLUSIONS.....	189
9.	Conculsion.....	190
10.	References.....	191
11.	ANNEX 1: GLOSSARY OF TERMS	198
12.	ANNEX 2: LEGAL REQUIREMENTS.....	202
12.1.	HEDF	202

TRUSTEE D3.1 TRUSTEE Core & Fair Knowledge

12.1.	HEDF	202
12.2.	AlmAAS	209
12.3.	TAI-SDF	223
12.4.	ATR	230
12.5.	KR.....	242
12.6.	AM.....	253
13.	ANNEX 3: SOCIO-ETHICAL REQUIREMENTS	266
14.	ANNEX 4: TAI-SDF Questionnaire for the assessment of the compliance with trustworthiness requirements 267	
15.	ANNEX 5: GDPR Compliance Checklist for TAI-SDF and STM.....	307

Index of figures

Figure 1: TRUSTEE Detailed architecture. WP3 and WP4 ABBs.....	18
Figure 2: Relationship between D3.1 and other WPs and related deliverables	20
Figure 3 HEDF Architecture	37
Figure 4 Semantic Representation layer	38
Figure 5 Low-level architecture of Communication Management Layer data flow.....	39
Figure 6 HEDF encryption layer architecture	40
Figure 7 HEDF encryption layer operation flow in a sequence diagram.....	42
Figure 8 HEDF encryption layer operation flow in the architecture	43
Figure 9 HEDF encryption layer first implementation functional components.....	45
Figure 10 HEDF middleware layer sequence diagram	48
Figure 11 HEDF Encryption Layer first implementation operation flow	49
Figure 12 HEDF encryption layer first implementation details.....	50
Figure 13 HEDF encryption layer first implementation functional test 1	54
Figure 14 HEDF encryption layer first implementation functional test 2	54
Figure 15 HEDF encryption layer first implementation functional test 3	55
Figure 16 HEDF encryption layer first implementation functional test 4	55
Figure 17 HEDF encryption layer first implementation functional test 5	56
Figure 18 HEDF encryption layer first implementation functional test 6	56
Figure 19 HEDF encryption layer first implementation functional test 7	57
Figure 20 HEDF encryption layer first implementation functional test 8	57
Figure 21 HEDF encryption layer first implementation functional test 9	58
Figure 22 HEDF encryption layer first implementation functional test 10	58
Figure 23: TRUSTEE AIaaS general component functionality and purpose	66
Figure 24: AIMaaS component diagram.....	71
Figure 25. The main interfaces and functionalities of the AIMaaS/FL module	73
Figure 26: xAI subcomponent Inputs/Outputs.....	74
Figure 27: Model Provider Sequence diagram.....	76
Figure 28 Prosumer use case.....	77

Figure 29 Prosumer use case UI (<i>mockups</i>).....	78
Figure 30. Initialization steps of the FL procedure	80
Figure 31. Steps for performing a homomorphically encrypted FL round.	81
Figure 32: AIaaS TRUSTEE canvas.....	84
Figure 33 Example of docker-compose.yaml file. It starts several containers including a Carla simulator, a visualization tool and a SLAM algorithm.	85
Figure 34 Example of the visualization tool	86
Figure 35 Impact of FL on improving raw data quality using DU models and the benefits on Lidar SLAM.....	93
Figure 36. Impact of FL on improving DL-based feature extraction on SLAMs	94
Figure 37. A Python implementation of the homomorphic encryption process that is adopted	95
Figure 38. L1 loss of the derived global model from the proposed deep unrolling FL scheme with and without homomorphic encryption vs communication rounds along with the best accuracy achieved by the centralized deep unrolling model.	96
Figure 39: SHAP Feature Importance Plot	97
Figure 40: SHAP Summary Plot	98
Figure 41: SHAP Dependence Plot.....	99
Figure 42: SHAP Dependence with Interaction Plot	99
Figure 43: SHAP Force Plot	99
Figure 44: SHAP explanations clustered by similarity Plot.....	100
Figure 45: SHAP Waterfall Plot	100
Figure 46. Given a low-resolution 3D point cloud derived from a 16-channel lidar, it is project onto a 2D low-resolution range image. This image is provided as input to the proposed deep unrolling model to estimate the corresponding high-resolution range image. The estimated image is then transformed into 3D coordinates, thus producing the desired high resolution point cloud for the lidar odometry problem.	101
Figure 47. The proposed end-to-end Deep unrolling model.	102
Figure 48. Impact of Explainable AI DU-model on improving raw data quality and the benefits on LiDAR SLAM	103
Figure 49: AI system lifecycle vs actors per each stage	107
Figure 50: TAI-SDF overall architecture	112
Figure 51: Sequence diagram related to the interaction between TAI-SDF and AIMaaS.....	114
Figure 52: Sequence diagram showing the process of TAI-SDF user authentication	115

TRUSTEE D3.1 TRUSTEE Core & Fair Knowledge

Figure 53: Sequence diagram showing the process of recording activities on ATR , such as add answers to questions, modify answers to questions	116
Figure 54: Sequence diagram related to interaction between GDPR compliance checker (inner module of TAI-SDF) and STM.....	116
Figure 55: screen related to GDPR compliance checker filling in questionnaire	117
Figure 56 High level architecture of ATR established on permissioned blockchain with the monitoring system.	128
Figure 57 The transaction recording phases at ATR.....	129
Figure 58 Sequence diagram showing the interactions to ATR.....	133
Figure 59 creation of single channel and network services running on docker.	135
Figure 60 Committing of chaincode on the peer nodes.	135
Figure 61 Transaction recording and query to the ledger	136
Figure 62 Endorsement of transaction failure due to duplicate transaction.....	136
Figure 63 Query to list all transactions recorded by TRUSTEE ABBs.....	137
Figure 64 Versioning method of KR referring to previous version of the knowledge content.	143
Figure 65 Automated Analysis using pandas profiling.....	144
Figure 66 High level architecture of KR enabling user to store content to the repository.	147
Figure 67 High level overview of Knowledge Graph technologies.....	148
Figure 68 Inner functioning of KR.....	149
Figure 69 TRUSTEE KR allowing user to interact to store knowledge.	150
Figure 70 The interaction to KR for Knowledge storage by authorised TRUSTEE entity.	153
Figure 71 Interaction from an ABB for the storage of knowledge to KR and the recording of CID to ATR.	154
Figure 72 Content stored to KR.	156
Figure 73 Interacting with KR to retrieve whole content.	156
Figure 74 Partial retrieval of content.	157
<i>Figure 75 Pinned content and removal of pinning for execution of garbage collection.</i>	<i>157</i>
Figure 76 Garbage collection to delete unpinned content.....	158
Figure 77 IPFS desktop integration to KR to avail IPFS services.	159
Figure 78: Overview of the 3 main entities of SSL.....	161
Figure 79: High-level architecture of the Authentication Manager interaction with other TRUSTEE related components.	165

TRUSTEE D3.1 TRUSTEE Core & Fair Knowledge

Figure 80: The external APIs and their protocols as exposed by the Authentication Manager.	166
Figure 81: Detailed architecture of the Authentication Manager showing all the internal components.....	167
Figure 82: Authentication Manager extension for GAIA-X users context.	169
Figure 83: Authentication Manager extension for GAIA-X users architecture.	171
Figure 84: Example showing the step-by-step process of authenticating a user, showing the use of both OIDC and SSI authentication.	179
Figure 85: Example showing the step-by-step process of onboarding a user, showing the use of both SSI and eIDAS.	180
Figure 86 Block diagram illustrating the functionalities involved in registering OIDC applications.	181
Figure 87 OIDC Authentication Process.....	182
Figure 88 Secure and verifiable user identity within the TRUSTEE ecosystem	183

Index of tables

Table 1: HEDF – User-centred requirements.....	32
Table 2: HEDF – Functional and non-functional technical requirements	35
Table 3 HEDF Architecture layers.....	36
Table 4 APIs for Communication Management within HEDF	39
Table 5 HEDF encryption layer first implementation functionalities.....	45
Table 6 HEDF encryption layer first implementation HE algorithms	46
Table 7 HEDF middleware layers technical details.....	52
Table 8 HEDF encryption layer first implementation technical details.....	53
Table 9: AIaaS – User-centred requirements.....	68
Table 10: AIaaS – Functional and non-functional technical requirements.....	69
Table 11: xAI subcomponent technical details	92
Table 12: TAI-SDF – User-centred requirements.....	108
Table 13: TAI-SDF – Functional and non-functional technical requirements	110
Table 14: ATR – User-centred requirements	126
Table 15: ATR – Functional and non-functional technical requirements.....	126
Table 16 Technical details for the first phase of implementation.	133
Table 17: KR – User-centred requirements	145
Table 18: KR – Functional and non-functional technical requirements	145
Table 19 The core functionalities of first implementation.....	152
Table 20 Technical details of first implementation phase	154
Table 21 Structure for knowledge storing to KR.....	155
Table 22: AM – User-centred requirements.....	162
Table 23: AM – Functional and non-functional technical requirements	164
Table 24: External interface between the Authentication Manager and multiple other ABBs to be used for user authentication.	171
Table 25: External interface between the Authentication Manager and the Accountable Transactions Recorder.	172
Table 26: External interface between the Authentication Manager and external SSI Agents (e.g. a user’s wallet).	173

TRUSTEE D3.1 TRUSTEE Core & Fair Knowledge

Table 27: External interface between the Authentication Manager and the eIDAS network.	174
Table 28: TAI-SDF – Legal requirements for system developers	223
Table 29: TAI-SDF – Legal requirements for end-users	224

LIST OF ABBREVIATIONS AND ACRONYMS

Acronym	Meaning
ABB	Architecture Building Block
AI	Artificial Intelligence
AIMaaS	AI Models as a Service
AM	Authentication Manager
AMQP	Advanced Message Queuing Protocol
API	Application Program Interface
ATR	Accountable Transaction Recorder
CA	Certificate Authority
CC	Crypto Context
CLI	Command-Line Interface
CNIL	Commission Nationale de l'Informatique et des Libertés (French Data Protection Authority)
DA	TRUSTEE Dashboard
DAPS	Dynamic Attributes Provisioning Service
DFS	Depth First Search
DID	Decentralized Identifiers
DL	Data Lake
DOI	Digital Object Identifier
DPIA	Data Protection Impact Assessments
DPO	Data Protection Officer
DSIL	Data Spaces Innovation Lab
DSILC	Data Spaces Innovation Lab Connector
DSS	Digital Signature Services
EAA	European Accessibility Act
EGI	European Grid Infrastructure
eIDAS	Electronic Identification and Trust Services
EOSC	European Open Science Cloud
EU	European Union
FAIR	Findable, Accessible, Interoperable, and Reusable
FAQ	Frequently Asked Questions
FHE	Fully Homomorphic Encryption
FLT	Fermat's Little Theorem
GDPR	General Data Protection Regulation
GraphDB	Graph Database
GUI	Graphical User Interface
HDFS	Hadoop Distributed File System
HE	Homomorphic Encryption
HEDF	Homomorphic Enabled Data Fusion
HLF	Hyperledger Fabric
HTTP	Hypertext Transfer Protocol
ICT	Information and Communication Technologies
IDS	International Data Spaces
IDSA	International Data Spaces Association
IPFS	InterPlanetary File System
IoT	Internet of Things
JSON	JavaScript Object Notation

TRUSTEE D3.1 TRUSTEE Core & Fair Knowledge

KPI	Key Performance Indicators
KR	Knowledge Repository
ML	Machine Learning
MT	Monitoring Tool
NIST	National Institute of Standards and Technology
NOSQL	Not Only Structured Query Language
OIDC	OpenID Connect
OneSS	One-Stop-Shop
OPC-UA	Open Platform Communication Unified Architecture
ORC	Optimized Row Columnar
OWL	Ontology Web Language
PACTS	Privacy and security Assessment for Cross-discipline Trust and Self-Sovereignty
ParIS	Participant Information Service
PIA	Protection Impact Assessments
pk	Public key
QA	Quality Assurance
RAM	Reference Architecture Model
RDF	Resource Description Framework
REST	Representational State Transfer
RPA	Robotic Process Automation
SDK	Software Development Kit
sk	Secret key
SPARQL	Protocol and RDF Query Language
SSI	Self-Sovereign Identity
SSI-HE	SSI Homomorphic Capable Framework
STAC	Spatio-Temporal Asset Catalog
STM	Security and Trust Manager
SW	Software
TAI-SDF	Trustworthiness Artificial Intelligence Support Design Framework
TBD	Trunk-based development
UI	User Interface
URL	Uniform Resource Locator
UX	User Experience
VA	Virtual Assistant
VC	Verifiable Credential
VDR	Verifiable Data Registry
VO	Virtual Organization
WP	Work Package
ZKP	Zero Knowledge Proof

1. EXECUTIVE SUMMARY

TRUSTEE acknowledges that today consumers must be very careful about sharing/accessing data, and national and international regulators/legislators step up privacy requirements. The industry is in need to identify how investing in data protection and privacy can create a business advantage for the proper handling of data, consent, notice, and regulatory obligations; how data is securely shared with third parties; how data is legally collected or stored; how data can be collected for processing; and regulatory restrictions such as GDPR, HIPAA, GLBA, or CCPA. Several effective actions have emerged seeking to address enhanced consumer-privacy and data-protection requirements. These span the life cycle of enterprise data, and expand the processes of operations, infrastructure, and customer-facing practices, and are enabled by data mapping.

This document, D3.1 – “Live document, TRUSTEE Automation, OPEN CONTINUUM & SSI Homomorphic capable framework” is the first deliverable of Work Package (WP) 3 – “TRUSTEE Core & Fair Knowledge”, in which six different TRUSTEE Architectural Building Blocks (ABBs) are discussed, as they were defined in D2.1 – “Live doc conceptualisation, use cases and system architecture V1” and as they are developed following the approaches regarding the “TRUSTEE DEVOPS AND CI/CD METHODOLOGY” as defined in TRUSTEE D4.1 TRUSTEE Automation, OPEN CONTINUUM & SSI Homomorphic capable framework: ANNEX 5: DEVOPS AND CI/CD METHODOLOGY.

- TRUSTEE Homomorphic Enabled Data Fusion (HEDF).
- TRUSTEE AI Model As A Service (AIaaS)
- TRUSTEE Trustworthiness AI Support Design Framework (TAI-SDF)
- TRUSTEE Accountable Transaction Recorder (ATR).
- TRUSTEE Knowledge Repository (KR).
- TRUSTEE Authentication Manager (AM).

The current deliverable defines the ABBs current design and implementation. As this deliverable is a living document, it is intended to serve as a real-time representation of the ongoing status of WP3. Consequently, the subsequent versions of the deliverable will provide updates and enhancements regarding the ABB descriptions and will continue to discuss the ongoing implementation activities and their outputs across the lifecycle of the project.

2. INTRODUCTION

In this document we follow the architectural designs of TRUSTEE's, a data-driven platform aimed at fostering reliable and eco-conscious management and processing of data. In the subsequent chapters, we outline our core and knowledge architectural blueprints with the intention of fulfilling the needs of the populace (referred to as consumers), the industrial sector, and the research community.

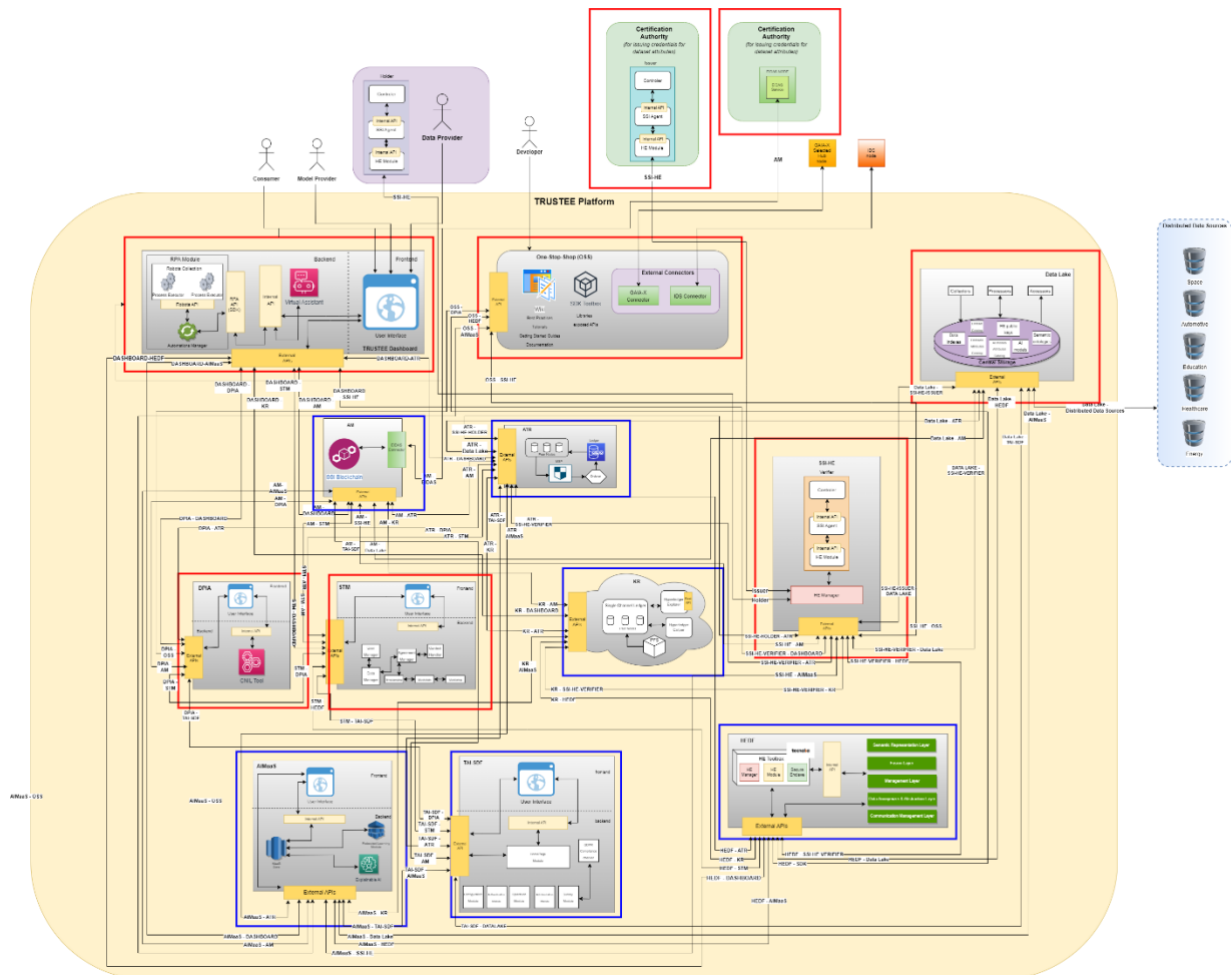


Figure 1: TRUSTEE Detailed architecture. WP3 and WP4 ABBs

This involves the facilitation of federated and FAIR data access within a discreet structure composed of both public and private entities. Our design is geared towards effecting FAIR communication with open data repositories and platforms such as GAIA-X, EOSC, and EGI. This communication involves the reciprocal exchange of current and future data sourced from diverse EU data domains, all of which is guided by the principles of open science and the established policies and procedures within the European Union. TRUSTEE stakeholders are endowed with user-friendly, secure, dependable, transparent, traceable, and environmentally sustainable ICT services. These services cater to data compilation, retention, manipulation, retrieval, and distribution. Our approach integrates state-of-the-art technologies to expedite the secure sharing and processing of data, a prime example being homomorphic encryption in adherence to pertinent EU regulations like the General Data Protection Regulation (GDPR). This extends to both data processors and subjects, along with other vested parties. A paramount focus is placed on societal innovation and privacy impact assessment via TRUSTEE technologies. This entails optimizing data processing at the periphery, bolstering resilience, managing transmission and storage, and refraining from superfluous manipulations, all in alignment with principles governing dependable and ethical AI.

2.1. PURPOSE AND SCOPE

This deliverable (D3.1 TRUSTEE CORE & FAIR KNOWLEDGE) is a live document including all the details, completed and ongoing work within WP3. More specifically we focus on the following within the context of the TRUSTEE CORE and FAIR Knowledge framework. This involves several key objectives:

- Firstly, the framework will facilitate the seamless integration of diverse AI models into the data processing pipeline. This functionality empowers users to customize or extend existing AI models to cater to specific needs.
- Moreover, the development includes the establishment of a Compliance Support Tool. This tool will furnish a standardized protocol for evaluating the alignment of the target object with the defined parameters. To further support technology providers, the framework will incorporate mechanisms for designing solutions that adhere to ethical and robust principles.
- In addition, a pivotal component of the initiative is the creation of the TRUSTEE Accountable Transactions Recorder (ATR), which will play a key role in ensuring transparency and accountability. The initiative will also encompass the construction of a cutting-edge FAIR Knowledge Repository. This repository will be dynamically informed by insights from experts, seamlessly integrated into the research hierarchy for effective knowledge dissemination.
- An essential advancement involves the conceptualization and implementation of a Self-Sovereign Identity concept, utilizing the eIDAS Platform as a foundation.
- Furthermore, an Innovative Explainable AI (xAI) tool is being developed. This tool aims to simplify the utilization of AI models as services, while maintaining a high degree of explainability.
- Lastly, the framework will integrate efficient data harmonization techniques, particularly concerning unit conversion, to ensure the coherence of heterogeneous data sources.

2.2. RELATION WITH OTHER WORK PACKAGES AND DELIVERABLES

This document focuses on the design and development of the “Knowledge repository layer” that is responsible for fusing the data coming from those heterogeneous sources and will be integrated based on data homogenization and semantic alignment toolkits that will be researched and developed, enabling effective cross-domain data search, retrieval and sharing. Alongside, direct integration and data calibration of related infrastructures’ decision-making tools will and decision support capabilities will be laid out to the public, stakeholders, and policymakers. Figure 1 represents the detailed architecture of TRUSTEE as presented in D2.1. For a better understanding of the document, ABBs remarked in blue correspond to the ones described in this document and the ones in red are detailed in the D4.1.

Figure 2 represents the relationship between this deliverable and the work done in other WPs in TRUSTEE. D3.1 had received input from WP2 in terms of background, user requirements, social and ethical requirements, and architectural principles that were taken in to account for designing the TRUSTEE Core and FAIR knowledge layers while it provides output in terms of technical requirements and technical implementation descriptions towards realizing the overall TRUSTEE platform. In particular, D3.1 provides output to WP5 through the generation of software (SW) modules and receives input in terms of data and feedback for further updates. Finally, WP3 and WP4 are highly interrelated as each of them presents and develops different components that are part of TRUSTEE which interact with each other.

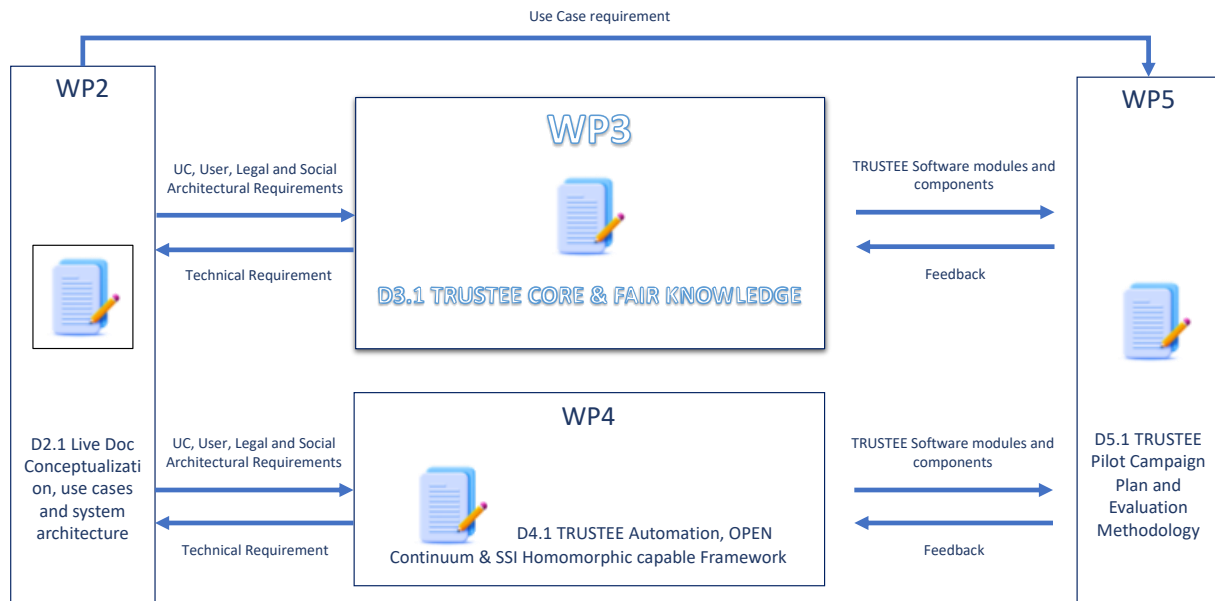


Figure 2: Relationship between D3.1 and other WPs and related deliverables

2.3. APPROACH AND STRUCTURE

This document described the work performed within WP3 that has been organized our workplan in 6 tasks running in parallel, that we briefly present below:

- **Task 3.1 TRUSTEE Homomorphic enabled data fusion** [Leader: ACCELI]. This task will focus on realizing the part of the TRUSTEE framework that is intended for data fusion. This will include acquisition of the data from all data sources, utilization of information from multiple sources, data abstraction, data management models, data minimization, fast cross-sectorial data query without violating IPRs, and deployment of a unified indexing framework ensuring system-level scalability, responsiveness, and reliability. Automatic processes will be developed to keep all the collected data within TRUSTEE Data Lake.
- **Task 3.2 AI model as a Service** [Leader: FORTH]. This task will enable on-boarding of various AI models for data processing providing the ability to tailor or even extend current AI running over TRUSTEE. The task will investigate how multiplexing and use of various AI models gives the ability to have different paradigms running for different data sets acquired by TRUSTEE e.g. federated learning for small fast analytics need in the source while on the same time to be able to run DNN in large multi-disciplinary request in multi-source queries will be considering. Enabling plug-n-play of AI models to the queries upon retrieval will result to better utilisation of the Data sources.
- **Task 3.3 Certification and auditing** [Leader: RINA]. This task will enable the certification and auditing for enabling access to the security-agnostic stakeholders that want to utilise TRUSTEE. To achieve this a Compliance Support Tool (CST) will be built to provide a standardised programme for assessing the conformance of the target object in scope (i.e. medical device, information system, management systems, etc.) with standards and it supports the user in conducting an assessment of the product/system development process during all the phases of its lifecycle in order to set up an effective internal control system focused on managing significant risks and to verify the compliance of the whole process to the relevant standards. Next, CST targets the Ethical AI assessment of AI technologies with the goal to support technology providers in designing solutions respecting the ethics and robustness principles. This task will also build TRUSTEE Accountable Transactions Recorder (ATR) that will utilise different technological components for recording every single transaction of the users with the data while

permitting private transactions where the content of the transaction can only be visible to authorized users as in the distributed ledger there are no administrative privileges. In this effort, it can also avail privacy-preserving features, such as anonymity and unlikability. Finally, this auditing process will be integrated for providing a culture of trust among the relevant stakeholders as they are certified and audit to use the TRUSTEE Facilitator.

- **Task 3.4 TRUSTEE Knowledge Repository [Leader: ENU].** Throughout its exploitation, TRUSTEE will generate large quantities of processed data, as results of detection events, feedback, crawling and monitoring as well as automated analysis. The knowledge repository aims to store search query results and computation results upon the informed consent of the users (e.g., data providers providing the datasets and data consumers performing the search queries and computations). One subcomponent of the Open Knowledge Repository follows the FAIR model. The knowledge graph aggregates information about the structure of the data itself, relational information of its sources and social connectivity. Additionally, TRUSTEE, proposes the collaboration among well recognized experts in the area, thus building a state-of-the-art FAIR capable knowledge repository, based on live feedback from the experts and integrated in the research chain of commands in knowledge sharing making it expert-friendly and subject-friendly while at the same time keep it as simple as possible.
- **Task 3.5 EIDAS, Authentication service, Sensor Enabler [Leader: INQBIT].** This task will build upon the eIDAS trust framework and will support the development and implementation of the Self-Sovereign Identity concept over the eIDAS Platform. The main objective is to showcase the deployment of blockchain technology that could leverage and benefit from the eIDAS regulatory framework. Additionally, this task will enable the onboarding of devices in a trusted and secure way while also guarantee the identification of the stakeholders that are utilising TRUSTEE.

Task 3.6 Explainable AI functions [Leader: K3Y]. This task aims to build an Innovative Explainable and Trustworthy eXplainable AI (xAI) tool, designed to streamline the utilization of AI models as a service. The primary objective is to address a prevailing challenge—the underutilization of data, often collected and stored for specific use cases but left dormant thereafter. The TRUSTEE xAI tool is poised to revolutionize this landscape by deploying customized techniques that incorporate advanced deep learning mechanisms. Its overarching goal is to bolster existing automated stakeholder summary techniques, providing user-centric and more comprehensive reports. To achieve this, the project will integrate cutting-edge data harmonization strategies. This involves optimizing data alignment through unit conversion, harmonizing preferred terminologies, and enabling seamless translation. By enhancing the efficiency and utility of data, these harmonization schemes are set to significantly benefit TRUSTEE stakeholders, furthering the initiative's overarching objectives.

The output of the above tasks is reported in this deliverable. This document presents the mapping of each WP3 task to each part of the deliverable.

Part I refers to Task 3.1 TRUSTEE Homomorphic enabled data fusion that described our efforts towards the integration and harmonization of data from various sources and domains by creating and implementing a Homomorphic Enabled Data Fusion (HEDF) component within the TRUSTEE ecosystem. This involves utilizing information from multiple sources to minimize the impact of individual source-generated noise and to fill in data gaps. This part is dedicated to the HOMOMORPHIC ENABLED DATA FUSION (HEDF) ABB

Part II refers to Task 3.2 AI model as a Service and Task 3.6 Explainable AI functions that its goal is to enable a plug-and-play method in the TRUSTEE ecosystem, emphasizing the smooth integration and customization of AI models to improve data processing capabilities. A key focus of this initiative within TRUSTEE is to simplify the onboarding process for various AI models, empowering TRUSTEE to enhance and expand its current AI infrastructure hosted in the cloud. This part is dedicated to AI Model As A Service, Explainable AI and Federated Learning, ABBs and submodules of AIaaS ABB.

Part III refers to Trustworthiness AI Support Design Framework (TAI-SDF) and **Part IV** refers to ACCOUNTABLE TRANSACTION RECORDER (ATR), that aim to define functional and technical components, specifying user requirements, determining system specifications, outlining the overall architecture, and marking out the legal requirements of tools developed in the scope of T3.3. Within T3.3 “Certification and Auditing”. This part is dedicated to Trustworthiness Artificial Intelligence Support Design Framework (TAI-SDF) and the latter one is named Accountable Transactions Recorder (ATR) ABBs.

Part V refer to Task 3.4 TRUSTEE Knowledge Repository aiming towards technologies required to develop, implement, and utilise a distributed peer-to-peer content-sharing protocol. This task is dedicated to Knowledge Repository (KR) ABB.

Part VI refers to Task 3.5 EIDAS, Authentication service, Sensor Enabler, with the objective to develop, implement and utilise a Self-Sovereign Identity (SSI) based on blockchain technology to allow to authenticate TRUSTEE users and bridging our system with the eIDAS network for retrieving a user’s identity information from a trusted source. This part is dedicated to the AUTHENTICATION MANAGER (AM) ABB.

For each chapter (for each ABB) the same structure has been followed:

- Introduction. This is the introductory section of the ABB, in which the needs and main objectives of the ABB are presented. Additionally, the main innovation is identified.
- Background. This section provides some basic ideas about the technologies considered by the ABB. The idea is to provide the necessary details to better understand the ABB architecture, operation and implementation. The detailed explanation of all the considered technologies is gathered in D2.1 (TRUSTEE consortium, 2023).
- Requirements. This section analyses the current compliance level of each ABB against TRUSTEE user-centred, technical, and legal and socio-ethical requirements defined in D2.1 (TRUSTEE consortium, 2023).
- Architecture. This section presents the design (high level and low level) of the ABB. The architecture, as well as the operational flows are detailed.
- Implementation. This section presents the current implementation level of the ABB at the time of writing in terms of functionalities, operation, and technical details.
- Validation. This section introduces any initial validation tests that have been carried out considering the maturity of the respective ABB.
- Next Steps. This section presents the next steps to be followed for the development of each ABB, which will be further discussed in upcoming versions of this deliverable.
- Conclusions. This section presents a summary of the objective and current status of the ABB.

Furthermore, the deliverable includes the following annexes:

- Annex 1: Glossary of Terms, a list of relevant terms with their definitions.
- Annex 2: Legal Requirements, gathering the preliminary compliance status with legal requirements gathered in D2.1 (TRUSTEE consortium, 2023) for each considered ABB. This information has been included as annex due to its long extension.
- Annex 3: Socio-ethical Requirements, gathering the preliminary compliance status with socio-ethical requirements gathered in D2.1 (TRUSTEE consortium, 2023) for each considered ABB. This Annex will be mostly analysed in later version of the deliverable.
- Annex 4: TAI-SDF Questionnaire for the assessment of the compliance with trustworthiness requirements.
- Annex 5: GDPR Compliance Checklist for TAI-SDF and STM

3. PART I: HOMOMORPHIC ENABLED DATA FUSION (HEDF)

3.1. INTRODUCTION

3.1.1. NEED AND OBJECTIVES

The following chapter presents an introduction to the work addressing integration and harmonization of data from diverse sources and domains, through the design and development of a Homomorphic Enabled Data Fusion (HEDF) component within the TRUSTEE ecosystem. By leveraging information from multiple sources, the objective is to reduce noise levels caused by individual sources and to fill in gaps in the available data.

The chapter also describes the data abstraction approach, empowered by homomorphic encryption, which builds upon an existing data management model. This approach follows a multi-layered architecture, with each layer corresponding to different levels of granularity. The system will provide users with varying access levels according to their requests and roles, enabling efficient cross-sectorial data queries while upholding intellectual property rights.

In terms of data management, the chapter also describes the design process of a unified indexing framework to ensure the scalability, responsiveness, and reliability of the system. The objective is to leverage and enhance existing techniques for multi-disciplinary data fusion, extracting more informative features and minimizing the amount of stored information. Adhering to the principle of data minimization, the project aims to strike a balance between retaining valuable insights and respecting privacy and data protection regulations.

This chapter sets the stage for the subsequent work in realizing the data fusion component of the TRUSTEE framework.

3.1.2. INNOVATION

By employing advanced techniques and methodologies, the HEDF aims to address the challenges of integrating, harmonizing, and extracting valuable insights from diverse data sources while guaranteeing their confidentiality through the use of advanced encryption technologies that allow privacy protection from outsiders (third-parties), and novel encryption key management solutions to protect data among the data sources. As a result, HEDF ultimately contributes to the advancement of data fusion in a multi-disciplinary context guaranteeing interoperability and confidentiality at all levels.

3.2. BACKGROUND

3.2.1. DATA ACQUISITION

3.2.1.1. DEFINITION OF DATA ACQUISITION

Data acquisition plays a crucial role in the data fusion process by collecting information from diverse sources. In recent years, advancements in technologies such as IoT devices, cloud computing, and big data analytics have revolutionized data acquisition methods. Traditional data acquisition approaches focused on structured data from databases, but with the proliferation of unstructured and semi-structured data, new techniques have emerged. For instance, web scraping and crawling techniques are widely used to extract information from websites and online platforms. Additionally, sensor networks and mobile devices enable real-time data collection, providing valuable insights for various applications. With the increasing availability of open data initiatives, researchers and practitioners can access a wealth of information from governmental and public sources. These developments have

expanded the possibilities of data acquisition, enabling comprehensive and diverse datasets for data fusion processes (Mittal 2021).

The emergence of Gaia-X and Open Cloud platforms has further enriched the landscape of data acquisition. Gaia-X, a European initiative, aims to foster the creation of a federated and secure data infrastructure, enabling cross-sectorial data sharing and collaboration. It provides a framework for secure and transparent data exchange, ensuring data sovereignty and privacy. Open Cloud platforms, on the other hand, offer scalable and flexible data storage and processing capabilities. With their emphasis on open standards and interoperability, these platforms enable seamless integration of data from various sources. Leveraging these initiatives, researchers can tap into a vast array of data sources, ranging from industrial IoT networks to healthcare systems, fostering innovative data fusion techniques (Candelieri 2022).

The scientific community has also recognized the importance of data acquisition in advancing research endeavours. Scientific publications serve as a valuable source of information, providing insights across various disciplines. Researchers have developed techniques to extract structured data from unstructured scientific literature, enabling integration with other datasets. Natural Language Processing (NLP) and text mining methods are employed to extract relevant information such as entities, relationships, and concepts from scientific texts. This allows researchers to leverage the collective knowledge captured in scientific publications, enhancing the depth and breadth of data available for fusion. Incorporating scientific literature into data fusion processes offers the potential for interdisciplinary discoveries and facilitates the creation of comprehensive knowledge bases (Ge 2022).

3.2.1.2. ADVANTAGES OF UTILIZING INFORMATION FROM MULTIPLE SOURCES

Utilizing information from multiple sources in data fusion processes offers several distinct advantages. Firstly, it enables the reduction of noise levels created by individual sources. Each data source may have its own limitations, biases, and inaccuracies. By fusing data from multiple sources, the noise introduced by individual sources can be mitigated through techniques such as outlier detection, data cleansing, and statistical modelling. This leads to improved data quality and enhances the reliability of the fused data (Ahmad 2022).

Another advantage of leveraging information from multiple sources is the ability to fill in missing data. In real-world scenarios, certain data sources may have incomplete or missing data points due to various reasons such as sensor failures, communication errors, or data collection limitations. By combining data from different sources, it becomes possible to fill the gaps and reconstruct missing information. This can be achieved through techniques like data imputation, interpolation, or leveraging correlations and patterns across sources. By completing the missing data, a more comprehensive and accurate representation of the underlying phenomena can be obtained (Singh 2021).

The multidisciplinary nature of data fusion necessitates the utilization of multiple sources to capture diverse perspectives and insights. Different data sources can provide complementary information that, when combined, leads to a more holistic understanding of complex phenomena. For example, combining data from IoT devices, social media, and satellite imagery can provide a comprehensive view of urban dynamics, including traffic patterns, social behaviour, and environmental conditions. This multidimensional view allows for more informed decision-making, better predictions, and improved situational awareness across various domains such as smart cities, healthcare, and environmental monitoring (Huang 2022).

3.2.2. DEFINITION OF DATA ABSTRACTION

Data abstraction is a fundamental concept in computer science and software engineering, providing a way to manage complexity and organize data by hiding implementation details and exposing only relevant information to users. It enables the creation of robust and maintainable software systems, facilitating modular design and

enhancing code reusability. In this article, we will delve into the definition and principles of data abstraction, exploring its significance and impact on modern computing.

Data abstraction provides a way to address the complexities inherent in data fusion. By creating abstract data models and interfaces, it enables data from disparate sources to be represented in a uniform and standardized manner. Here's how data abstraction facilitates data fusion (Zhang Z 2022):

Unified Data Representation: Abstraction allows developers to design data models that abstract away the underlying data formats and representations. By doing so, data from various sources can be converted into a common format, making it easier to integrate and process the information seamlessly.

Semantic Mapping and Mediation: Data abstraction enables the creation of semantic mappings and mediators. Semantic mapping aligns the meaning of data elements across different sources, ensuring that data is correctly interpreted and fused. Mediators act as intermediaries, translating data between different formats or protocols, facilitating interoperability between data sources.

Data Quality Management: Data abstraction can incorporate mechanisms for data quality management. By defining quality attributes and metadata, the fusion process can account for data reliability, accuracy, completeness, and timeliness, among other aspects. This ensures that only high-quality data contributes to the fused dataset.

Hierarchical Abstraction: In scenarios where data sources have varying levels of granularity, hierarchical abstraction allows the fusion process to handle data at different levels of detail. This flexibility ensures that the fusion process captures essential information without losing granularity where it is crucial.

Contextual Abstraction: Data fusion often involves considering data in different contexts. Data abstraction can capture context-specific information, making it possible to interpret and fuse data appropriately within specific use cases or domains.

Data abstraction also plays a vital role in enabling data interoperability in the context of data fusion (Guizzardi 2023):

Standardized Interfaces: Abstraction allows the creation of standardized interfaces and APIs for accessing the fused data. These interfaces hide the underlying complexity and data representation, enabling different applications to access the data without needing to understand the specifics of the fusion process.

Semantic Consistency: Data abstraction helps ensure semantic consistency across different systems by providing a unified view of the data. This ensures that data is interpreted correctly, regardless of the application using it.

Modularity and Extensibility: Abstraction promotes modularity and extensibility in the design of fusion systems. By encapsulating data fusion logic within well-defined modules, interoperability with new or updated components becomes more manageable.

Data Access Control: Data abstraction can incorporate access control mechanisms to manage data privacy and security. This ensures that only authorized applications or users can access specific data elements within the fused dataset.

Data Transformation: Similar to data fusion, data abstraction facilitates data transformation, allowing data to be translated into formats compatible with different applications or systems.

3.2.2.1. BUILDING ON AN EXISTING DATA MANAGEMENT MODEL

Building on an existing data management model is a common practice in the field of data engineering and software development. Leveraging an established model allows saving time and effort while benefiting from the best practices and solutions already implemented. Below, we present the state-of-the-art process for building on an existing data management model.

1. Understanding the Existing Data Management Model: Before building on an existing model, it's crucial to gain a comprehensive understanding of the model's structure, components, and functionalities. Analyze the data schema, storage mechanisms, indexing strategies, access patterns, and any constraints or limitations. This step ensures that we know the strengths and weaknesses of the model, enabling us to make informed decisions when extending or modifying it. Within the context of TRUSTEE, we study the TRUSTEE architecture and the data models available in order to determine which features should be taken into consideration for the unified indexing framework.

2. Identifying Requirements and Gaps: Assess the specific requirements of your project and identify any gaps in the existing data management model. Determine what functionalities or features are missing and need to be added to support your application's needs. This step involves gathering input from stakeholders, domain experts, and end-users to ensure that the extended model aligns with the project's objectives. Within TRUSTEE, the user-centred requirements that have been identified and documented within D2.1 will be taken into consideration for the design and development of the work within T3.1 and the HEDF ABB.

3. Extending the Data Model: Based on the identified requirements, we can extend the existing data management model to accommodate new features or data structures. This may involve introducing additional tables, modifying existing schemas, or incorporating new data types. We should also ensure that the extensions are compatible with the existing model and do not disrupt its integrity or performance. This and the following steps will be examined within TRUSTEE after the first pilot phase and the evaluation results that will stem from the data analysis.

4. Data Migration and Integration: If the extension involves changes to the data model's schema or storage format, a plan should be formulated for data migration and integration. Migrating data from the old model to the extended model while preserving data consistency and accuracy is crucial. This process should be well-tested and, if possible, performed incrementally to minimize downtime and risks.

5. Performance Optimization: The performance of the extended data management model should be monitored and optimized for efficient data retrieval, storage, and processing. Implement indexing, caching, and other performance-enhancing techniques to ensure that the system can handle increased data volume and access patterns.

6. Testing and Quality Assurance: Thoroughly testing of the extended data management model should be performed to validate its functionality, performance, and reliability. Unit tests, integration tests, and stress tests are suitable to identify and address any issues before deployment. Quality assurance is essential to ensure the extended model meets the expected standards and provides accurate and reliable data management.

7. Documentation and Knowledge Sharing: The documentation of the changes made to the existing data management model and provision comprehensive documentation for developers, maintainers, and future stakeholders. Proper documentation ensures that everyone involved understands the extended model and can work with it effectively.

3.2.3. DEFINITION OF A UNIFIED INDEXING FRAMEWORK

A unified indexing framework is a crucial component in the management and organization of data within the context of data fusion. It provides a systematic approach to indexing and querying data from diverse sources, enabling efficient retrieval and integration of information. A unified indexing framework establishes a common

structure and format for representing data, facilitating seamless interoperability between different data types and sources. This framework typically involves the use of standardized data models, schema mappings, and indexing techniques to ensure consistency and accessibility of data across the entire data fusion system (Liu W. 2022).

The adoption of a unified indexing framework offers numerous benefits in terms of scalability, responsiveness, and reliability. By leveraging a unified indexing structure, the system can efficiently handle large volumes of data and accommodate future expansions. The framework enables optimized data retrieval operations, ensuring that queries can be executed in a timely manner, even when dealing with massive datasets. Moreover, the indexing framework contributes to the reliability of the system by providing mechanisms for data redundancy, fault tolerance, and data consistency (Zhang L 2021).

With the increasing complexity and heterogeneity of data in data fusion scenarios, recent advancements have been made to enhance the capabilities of unified indexing frameworks. For instance, research has focused on the integration of semantic technologies, such as ontologies and knowledge graphs, into the indexing process. These semantic-based indexing frameworks enable a deeper understanding of data semantics, facilitating more intelligent and context-aware data retrieval. By incorporating semantic information, the indexing framework can capture relationships between entities, support complex query operations, and provide richer search capabilities (Li, Y. 2022).

Additionally, recent research has explored the utilization of distributed indexing frameworks to address the scalability challenges of data fusion systems. Distributed indexing frameworks leverage the power of distributed computing and parallel processing to enable efficient indexing and querying across a network of interconnected nodes. This approach allows for horizontal scaling, where the system can dynamically adapt to the increasing demands of data fusion applications. By distributing the indexing workload, the system can handle larger datasets, improve query response times, and enhance fault tolerance (He, J. 2021).

3.2.4.

IMPORTANCE OF SCALABILITY, RESPONSIVENESS, AND RELIABILITY

Data indexing, fusion, and interoperability are critical components of modern data-driven applications, enabling efficient data management and utilization. To ensure the success of these processes, three key factors play a pivotal role: scalability, responsiveness, and reliability.

With the ever-increasing volume of data being generated and processed, scalability is of utmost importance in data indexing, fusion, and interoperability. As datasets grow in size, the underlying systems must be able to handle the expanding workload without sacrificing performance. Scalability allows data management solutions to accommodate growing data demands, whether it involves indexing large datasets, fusing data from diverse sources, or supporting interoperability between various systems. Scalable solutions can adapt to changing requirements, making them future-proof and capable of handling the data challenges of tomorrow (Barbosa et al., 2019).

Additionally, in today's fast-paced digital landscape, responsiveness is critical for data indexing, fusion, and interoperability. Users and applications expect real-time access to data and seamless interactions. A responsive system ensures that data retrieval, indexing operations, and fusion processes are performed promptly, minimizing delays and providing timely insights. Whether it's retrieving specific information from an indexed dataset or seamlessly fusing data from different sources, responsiveness enhances user experience and operational efficiency.

The importance of reliability cannot be overstated in data indexing, fusion, and interoperability. Data accuracy and consistency are paramount, especially when dealing with vast datasets and critical applications. Reliable data indexing ensures that queries return accurate results, while reliable data fusion guarantees that information from multiple sources is correctly integrated. In the context of interoperability, reliable data exchange between different systems ensures seamless data flow and minimizes the risk of errors or data loss. Reliability builds trust in the data

management process and ensures the integrity of the insights and decisions derived from the data (Armbrust et al., 2021).

3.2.5. HOMOMORPHIC ENABLED DATA ENCRYPTION

The core component of the HEDF component is the homomorphic encryption layer, which provides confidentiality to the data transfer process between data providers and data consumers, as well as confidentiality to the processing operation over the data from data providers.

Moreover, data fusion from different data providers is feasible in this layer if the same homomorphic key is considered for the encryption process as only data encrypted with the same homomorphic encryption key can be processed together. For this purpose, as all the data providers participating on the fusion will use the same encryption key, symmetric cryptography is not possible as the decryption key will be the same, and all data providers participating on the fusion could decrypt other data providers data. For this reason, asymmetric cryptographic mechanisms have been considered to avoid data disclosure among participating data providers.

In TRUSTEE, two different homomorphic frameworks have been considered to cover different kind of operations to be confidentially executed. These are: OpenFHE¹ and TFHE² (details about these technologies are gathered in D2.1).

On the one hand, OpenFHE is the best option to support complex mathematical operations as it includes, for example, a Chebyshev approximation algorithm that allows to perform operations such as divisions, square roots or logistic functions. Furthermore, OpenFHE developer community is large, and they release very frequently. However, OpenFHE cannot deal with logic gates with public keys, as needed by comparisons or binary operations. That is why TFHE-rs is also considered, as it is the only HE option that allows this with public keys. More details of each HE framework can be found in D2.1.

3.2.5.1. HE KEYS GENERATION

The required HE keys are generated inside TRUSTEE platform to guarantee data confidentiality among data providers participating on the fusion as well as with data consumers. Following asymmetric cryptographic concepts, a public key will be considered for the encryption process while a private key will be considered for the decryption process. In TRUSTEE, the core “TRUSTEE platform” will be the only actor who knows the private key and is capable of decrypting.

- OpenFHE

OpenFHE cryptographic context is generated using one of the FHE schemes provided by the library, and from this context a private and a public key are generated. The public key and context need to be serialized as they need to be stored in a public repository, such as TRUSTEE Data Lake, so that they can be publicly accessible. Furthermore, the public key needs to be shared in such a way that it is accessible by all data providers participating on the fusion allowing all of them to homomorphically encrypt their data with the same key. However, as mentioned before, the private key is not shared with data providers to avoid data disclosure among them. Moreover, the context needs to be also shared with data providers for the encryption process as well as with data consumers as it is necessary to execute the algorithms over encrypted data.

¹ [OpenFHE: Open-Source Fully Homomorphic Encryption Library \(acm.org\)](https://openfhe.org/)

² [TFHE: Fast Fully Homomorphic Encryption Over the Torus | SpringerLink](https://www.springerlink.com/TFHE/)

- TFHE-rs

In TFHE-rs, however, there is no cryptographic context, but it is necessary to generate three different types of keys: public, private, and server keys. In the same way as in OpenFHE, the public key must be shared with the data providers to be able to encrypt their data while the server key is needed by data consumers for the algorithm's execution over encrypted data. As in the case of OpenFHE, the private key is not shared and remains inside TRUSTEE platform.

As it has been explained, keys are generated inside TRUSTEE platform. However, high levels of security are needed, especially for the private key generation, to prevent any actor with access to the TRUSTEE platform from accessing the private key and thus decrypting the data providers information. In this sense, the possibility of using a secure enclave (or similar device) has been analysed.

An important limitation of HE is the size of the keys. Public keys for both technologies are very large, especially in TFHE-rs that occupies up to 2.2 GB. This has been identified as a potential problem when keys need to be shared among the participants, but also for their generation inside a secure enclave (or similar device) as they are not usually prepared to manage these file sizes. Research work on how to solve these potential problems is ongoing and solutions will be presented in next versions of D3.1.

SECURE ENCLAVE

Secure enclaves, also known as Trusted Execution Environments (TEE), refer to a technique that isolates an environment (processor, memory and storage) with high levels of security limiting the exchange of information with other untrusted contexts. For example, a secure enclave at the hardware and operating system level could create and store private keys and perform operations with them such as information decryption without allowing the private keys to leave the secure enclave or to be loaded into memory of some untrusted process. A secure enclave provides a limited set of instructions to execute trusted operations in isolation from an untrusted application context.

The secure enclave needed in TRUSTEE fits in the definition of “cryptographic module” given by the EN ISO/IEC 19790:2020³ standard. This International Standard specifies the security requirements for a cryptographic module utilised within a security system protecting sensitive information in computer and telecommunication systems.

Security Levels

The standard defines four security levels to provide a wide spectrum of data sensitivity, and a diversity of application environments:

- **Security Level 1:** This level provides the security baseline. No specific physical security mechanisms are required in a Security Level 1 hardware cryptographic module beyond the basic requirement for production-grade components. Non-invasive mitigation methods or mitigation of other attacks which are implemented are documented. Such implementations are ideally appropriate for security applications where controls, such as physical security, network security, and administrative procedures are provided outside of the module but within the environment which it is to be deployed.
- **Security Level 2:** In this level, physical security mechanisms are enhanced, including tamper-evident seals or coatings. Role-based authentication is also required. Security Level 2 allows a software cryptographic module to be executed in a modifiable environment that implements role-based access controls or, at the minimum, a discretionary access control with robust mechanism of defining new groups

³ <https://standards.iteh.ai/catalog/standards/cen/543fdbea-52bf-4549-8a5a-d07f8b91625b/en-iso-iec-19790-2020>

and assigning restrictive permissions through access control lists, and with the capability of assigning each user to more than one group, and that protects against unauthorised execution, modification, and reading of cryptographic software.

- **Security Level 3:** This level provides additional requirements to mitigate the unauthorised access to sensitive security parameters (SSP). Security Level 3 also protects a cryptographic module against a security compromise due to environmental conditions outside of the module's normal operating ranges for voltage and temperature.
- **Security Level 4:** This security level provides the highest level of security defined in this standard. At Security Level 4, the physical security mechanisms provide a complete envelope of protection around the cryptographic module with the intent of detecting and responding to all unauthorised attempts at physical access when sensitive security parameters are contained in the module whether external power is applied or not. Penetration of the cryptographic module enclosure from any direction has a very high probability of being detected, resulting in the immediate zeroization of all unprotected SSPs. Security Level 4 introduces the multi-factor authentication requirement for operator authentication.

The overall security rating of a cryptographic module must be chosen to provide a level of security appropriate for the security requirements of the application and environment in which the module is to be utilised and for the security services that the module is to provide.

In the case of the secure enclave for TRUSTEE, Level 3 was selected for most of the requirements, with a few exceptions where Level 2 was chosen.

Types of Cryptographic Modules

A cryptographic boundary consists of an explicitly defined perimeter (i.e. set of hardware, software or firmware components) that establishes the boundary of all components of the cryptographic module. The requirements of the ISO/IEC 19790 apply to all algorithms, security functions, processes, and components within the module's cryptographic boundary. Non-security relevant algorithms, security functions, processes or components may be included within the cryptographic boundary.

According to the ISO/IEC 19790 standard, a cryptographic module should be defined as one of the following module types:

- **Hardware module** is a module whose cryptographic boundary is specified at a hardware perimeter. Firmware and/or software, which may also include an operating system, may be included within this hardware cryptographic boundary.
- **Software module** is a module whose cryptographic boundary delimits the software exclusive component(s) (may be one or multiple software components) that execute(s) in a modifiable operational environment. The computing platform and operating system of the operational environment which the software executes in are external to the defined software module boundary.
- **Firmware module** is a module whose cryptographic boundary delimits the firmware exclusive component(s) that execute(s) in a limited or non-modifiable operational environment. The computing platform and operating system of the operational environment which the firmware executes in are external to the defined firmware module boundary but explicitly bound to the firmware module.
- **Hybrid Software module** is a module whose cryptographic boundary delimits the composite of a software component and a disjoint hardware component (i.e. the software component is not contained within the hardware module boundary). The computing platform and operating system of the operational environment which the software executes in are external to the defined hybrid software module boundary.
- **Hybrid Firmware module** is a module whose cryptographic boundary delimits the composite of a firmware component and a disjoint hardware component (i.e. the firmware component is not contained within the hardware module boundary). The computing platform and operating system of the operational

environment which the firmware executes in are external to the defined hybrid firmware module boundary but explicitly bound to the hybrid firmware module.

The secure enclave to be developed fits in the description of hybrid software module, and therefore, all the requirements related to them apply.

Cryptographic Boundary

The cryptographic boundary of a **hybrid cryptographic module** shall:

- Be the composite of the module's hardware component boundary and the disjoint software or firmware component(s) boundary.
- Include the collection of all ports and interfaces from each component.

In addition to the disjoint software or firmware component(s), the hardware component can also include embedded software or firmware.

3.2.5.2. HE ENCRYPTION PROCESS

In TRUSTEE, data providers will carry out the HE encryption process on their own premises in order to share data once encrypted preventing their actors to access their data.

For both OpenFHE and TFHE-rs libraries, data providers will use the public key to encrypt their data and obtain ciphertexts. In the case of OpenFHE, it is also necessary to know the cryptographic context.

An important limitation in TFHE's ciphertext handling is that the encryption of numeric values is done by concatenation of shortint types. The shortints are 3-bit elements, that is, they can represent values between 0 and 7. TFHE offers the possibility of concatenating several of these objects by blocks of four at a time, which allows working with higher numerical ranges, but the computational cost grows as more blocks are concatenated.

As a result, data providers' data is homomorphically encrypted, guaranteeing confidentiality in their data transfer and processing.

3.2.5.3. OPERATIONS IN THE ENCRYPTED DOMAIN

As explained before, data consumers will be in charge of the encrypted (and fused) data processing. For this purpose, two different HE libraries have been considered to provide TRUSTEE with a large number of available algorithms to be executed on the homomorphic domain. Depending on the kind of operations needed by each algorithm, it is better to use OpenFHE or TFHE-rs. There are some operations that are only available in one of the libraries (theoretical limitations gathered in D2.1). However, although other operations can be considered in both of them, performance and complexity need to be analyzed to identify the most suitable framework in each case. Some examples of operations to be executed in TRUSTEE with OpenFHE are arithmetic operations, statistics, aggregations, and regressions. In the case of TFHE-rs, it is more suitable for searches and comparisons, for example.

3.2.5.4. HE DECRYPTION PROCESS

The decryption process uses the private key in both libraries as well. Additionally, in the case of OpenFHE the cryptographic context is also necessary. The decryption process only refers to the result of the algorithm applied on the homomorphic domain; in other words, encrypted data from data providers is never decrypted.

As mentioned before, the private key will not leave TRUSTEE platform; in fact, it will be securely generated and stored inside a secure enclave. For this reason, the decryption process, which requires the private key for its execution, will be also executed inside the secure enclave to avoid the private key exposure in untrusted contexts.

Once the decryption of the result is done, it will be shared again with the data consumer who has executed the specific algorithm.

3.3. REQUIREMENTS

In this chapter, the compliance of the HEDF ABB of TRUSTEE is investigated with regards to the user-centred, technical, legal, and socio-ethical requirements that were reported in D2.1 (TRUSTEE consortium, 2023). The results of this investigation provide further preliminary insight regarding with the aforementioned requirements addressed during the development and by the implementation of the DPIA and are presented in the following sub-sections. It is worth noting that only the Mandatory, Desirable, and Optional requirements have been considered at this point. Furthermore, only the currently compliant or partially-compliant requirements have been considered. As the implementation of the DPIA progresses during the development of the TRUSTEE Platform and the maturity of the ABB increases, the investigation of the compliance will evolve and further updates and/or enhancements will be provided in upcoming versions of this deliverable, namely D3.2 and D3.3.

User-centred requirements were identified through the UX Research that was conducted under T2.3. These requirements were reported in D2.1 (TRUSTEE consortium, 2023), while they were also utilized in order to extract technical, functional and non-functional requirements, that were also outlined in D2.1. Table 1 below presents the interplay between user-centred requirements and technical requirements, which is the functional and non-functional requirements that address each user-centred requirement. The user-centred requirement ID is used to reference the respective requirement reported in D2.1 (TRUSTEE consortium, 2023) while the requirement's description is also provided alongside whether the DPIA is currently fully or partially compliant with the user-centred requirement, i.e., indicating whether technical requirements have been identified in order to address the respective user-centred requirement. The ID of the technical requirement, as reported in D2.1 (TRUSTEE consortium, 2023), that is addressing each user-centred requirement is also provided in the last column of the table. Additionally, it is possible than more than one technical requirement addresses a single user-centred requirement.

Table 1: HEDF – User-centred requirements

Req id (WP2)	Req. Description	Compliance	How is addressed
USR-001	TRUSTEE must enable collaboration among domains and enable access and use of data for secondary use from multiple domains	Partially	Req-HEDF-FUNC-1
			Req-HEDF-FUNCT-2
			Req-HEDF-FUNCT-3
			Req-HEDF-FUNCT-4
			Req-HEDF-FUNCT-5
			Req HEDF -FUNCT-6
			Req-HEDF-FUNCT-7
			Req-HEDF-FUNCT-10
			Req-HEDF-FUNCT-11
USR-002	TRUSTEE must use standardized methods and clear instructions for correct data entry and formatting	Partially	Req-HEDF-FUNC-1
			Req-HEDF-FUNCT-2
			Req-HEDF-FUNCT-3
			Req-HEDF-FUNCT-4
			Req-HEDF-FUNCT-5
			Req HEDF -FUNCT-6
			Req-HEDF-FUNCT-7
			Req-HEDF-FUNCT-10

			Req-HEDF-FUNC-11
USR-003	Users must provide access to their data in a secure and privacy aware way without sending their raw data to others	Partially	Req-HEDF-FUNC-1 Req-HEDF-FUNCT-2 Req-HEDF-FUNCT-3 Req-HEDF-FUNCT-4 Req-HEDF-FUNCT-5 Req HEDF -FUNCT-6 Req-HEDF-FUNC-7 Req-HEDF-FUNC-9 Req-HEDF-FUNC-10 Req-HEDF-FUNC-11
USR-005	The data provided must be used with respect to privacy regulations, legal frameworks, and ethical considerations and without the risk of data loss	Partially	Req-HEDF-FUNCT-2 Req-HEDF-FUNC-7 Req-HEDF-FUNC-11
USR-009	TRUSTEE must protect education data from misuse that aims in profiling students	Partially	Req-HEDF-FUNCT-2 Req-HEDF-FUNC-7 Req-HEDF-FUNC-10 Req-HEDF-FUNC-11
USR-012	TRUSTEE should accept data in different formats with unequivocal labels	Yes	Req-HEDF-FUNC-1 Req-HEDF-FUNCT-3 Req-HEDF-FUNCT-4
USR-016	TRUSTEE should enable interoperability and integration of data from different data sources	Partially	Req-HEDF-FUNC-1 Req-HEDF-FUNCT-3 Req-HEDF-FUNCT-4 Req-HEDF-FUNCT-5 Req HEDF -FUNCT-6 Req-HEDF-FUNC-7 Req-HEDF-FUNC-10 Req-HEDF-FUNC-11
USR-018	TRUSTEE should be compatible with existing data exchange technology infrastructures, such as IDS and Gaia-X	Partially (in combination to T4.4)	Req-HEDF-FUNC-1 Req-HEDF-FUNCT-3 Req-HEDF-FUNCT-4 Req-HEDF-FUNCT-5 Req HEDF -FUNCT-6 Req-HEDF-FUNC-7 Req-HEDF-FUNC-10 Req-HEDF-FUNC-11
USR-029	TRUSTEE should offer collaboration across multi-disciplinary data domains	Partially	Req-HEDF-FUNC-1 Req-HEDF-FUNCT-2 Req-HEDF-FUNCT-3 Req-HEDF-FUNCT-4 Req-HEDF-FUNCT-5 Req HEDF -FUNCT-6 Req-HEDF-FUNC-7 Req-HEDF-FUNC-9 Req-HEDF-FUNC-10 Req-HEDF-FUNC-11

USR-032	TRUSTEE should enable query and align datasets with other data from different domains	Partially	Req-HEDF-FUNC-1 Req-HEDF-FUNCT-3 Req-HEDF-FUNCT-4 Req-HEDF-FUNCT-5 Req-HEDF-FUNC-7 Req-HEDF-FUNC-10 Req-HEDF-FUNC-11
USR-033	TRUSTEE should enable computations and analysis over the accessed data	Partially	Req-HEDF-FUNCT-2 Req-HEDF-FUNC-7 Req-HEDF-FUNC-10 Req-HEDF-FUNC-11
USR-035	TRUSTEE should support easy transform/reformat of data	Partially	Req-HEDF-FUNC-1 Req-HEDF-FUNCT-3 Req-HEDF-FUNCT-4 Req-HEDF-FUNCT-5 Req HEDF -FUNCT-6 Req-HEDF-FUNC-7 Req-HEDF-FUNC-10 Req-HEDF-FUNC-11
USR-036	TRUSTEE should allow efficient data querying	Partially	Req-HEDF-FUNCT-2 Req-HEDF-FUNC-7 Req-HEDF-FUNC-10 Req-HEDF-FUNC-11
USR-038	TRUSTEE should enable transfer of learning without IPR and data privacy issues	Partially	Req-HEDF-FUNCT-2 Req-HEDF-FUNC-7 Req-HEDF-FUNC-10 Req-HEDF-FUNC-11
USR-041	TRUSTEE should allow the process of data without IPR and privacy issues	Partially	Req-HEDF-FUNCT-2 Req-HEDF-FUNC-7 Req-HEDF-FUNC-10 Req-HEDF-FUNC-11
USR-042	TRUSTEE should be compliant with all security requirements and legislations regarding data manipulation and querying	Partially	Req-HEDF-FUNCT-2 Req-HEDF-FUNC-7 Req-HEDF-FUNC-10 Req-HEDF-FUNC-11
USR-045	TRUSTEE should securely share commercial and private data	Partially	Req-HEDF-FUNCT-2 Req-HEDF-FUNC-7 Req-HEDF-FUNC-10 Req-HEDF-FUNC-11
USR-046	TRUSTEE should eliminate the correlation of data with personal data	Partially	Req-HEDF-FUNCT-2 Req-HEDF-FUNC-7 Req-HEDF-FUNC-10 Req-HEDF-FUNC-11
USR-048	TRUSTEE should support the ability to automatically add data and include extra metadata fields to these data if needed	Partially	Req-HEDF-FUNC-1 Req-HEDF-FUNCT-3 Req-HEDF-FUNCT-4 Req-HEDF-FUNCT-5 Req HEDF -FUNCT-6 Req-HEDF-FUNC-7

		Req-HEDF-FUNC-10 Req-HEDF-FUNC-11
USR-050	TRUSTEE should utilize easy-to-use filters in data querying	Req-HEDF-FUNC-7 Req-HEDF-FUNC-10 Req-HEDF-FUNC-11
USR-064	TRUSTEE should enable users to provide their AI/ML algorithms for others to use them	Req-HEDF-FUNC-7 Req-HEDF-FUNC-9 Req-HEDF-FUNC-10 Req-HEDF-FUNC-11

Table 2 presents the preliminary results of the investigation regarding the compliance of the HEDF with the technical, functional, and non-functional requirements specified in D2.1. The ID of the respective technical requirement is used as a reference to D2.1, while the requirement's description is also provided alongside whether DA complies with the requirement and how. Further updates and/or enhancements regarding the approach presented below for addressing each technical requirement will be provided in upcoming versions of this deliverable, namely D3.2 and D3.3.

Table 2: HEDF – Functional and non-functional technical requirements

Req id (WP2)	Req. Description	Compliance	How is addressed
Req-HEDF-FUNC-1	TRUSTEE should be able to accept data from various sources and domains and fuse them based on common elements.	Partially	PPRL (privacy preserving record linkage) between the Health and the Education domains have been implemented. See Table 6 for more details. More general use cases will be defined in future versions of HEDF.
Req-HEDF-FUNCT-2	Data provided to TRUSTEE should always be homomorphically encrypted. Data will be processed in encrypted form.	Yes	The design and operation flow of HEDF requires data to be encrypted before being shared for computations.
Req-HEDF-FUNCT-3	TRUSTEE could be able to accept data in different formats without requiring the data provider to reformat them.	Yes	As the data resides in the data provider's premises, the TRUSTEE does not limit the types of formats for allowed data. However, the Data Provider is required to specify the formats of their data to provide clarity on the allowed operations for their data.
Req-HEDF-FUNCT-4	TRUSTEE should be able to catalogue and retrieve data from existing infrastructures that host data from various domains.	Partially	HEDF is working towards investigating data retrieval from different infrastructures.
Req-HEDF-FUNCT-5	TRUSTEE should be able to perform computations and analysis over the data selected by the user.	Yes	By design, HEDF requires information about the datasets to be considered for the computation from the SSI-HE based on the TRUSTEE data consumer requirements.
Req-HEDF-FUNCT-6	The TRUSTEE platform should be able to search across multiple data sources	Yes	By design, HEDF directly receives the specific sources to be considered from

	quickly and efficiently for the user selected criteria.		the SSI-HE based on the TRUSTEE data consumer requirements.
Req- HEDF-FUNCT-7	TRUSTEE should be able to perform simple aggregation of AI models in the frame of FL.	Yes	HEDF is able to support averaging and median fusion rules for aggregating AI models.
Req- HEDF-FUNCT-9	HE-enabled FL framework should be able to store the global NN model in an unencrypted fashion to the Data Lake	Yes	This is supported by the HEDF module

A preliminary analysis of the compliance with the legal and socio-ethical requirements defined in the D2.1 (TRUSTEE consortium, 2023) is shown in ANNEX 2 and ANNEX 3.

3.4. ARCHITECTURE

3.4.1. HIGH LEVEL

The HEDF architecture will consist of a multi-layered approach, with 3 main layers comprising layers comprising the functionalities that should be available for the proper data management and fusion, as shown in Table 3. These 3 layers encompass functionalities previously described as separate layers in D3.1, so that they can provide a more concise and robust view of the existing and future implementation of HEDF.

Table 3 HEDF Architecture layers

HEDF Layer	Description	Technical implementation
Semantic representation layer	Transform collected data into knowledge, by applying common/standardized semantic ontologies on the data, correlated with the appropriate contextual information	Semantic annotation model → RDF ontologies that represent TRUSTEE datasets and their concepts using open standards
Management layer	Categorize the data, to allow for better processing and comprehension from the TRUSTEE platform	Unified Indexing Framework → dataset indexes Communication with external data sources (e.g. GAIA-X, IDS connectors)
Encryption layer	Fuse different datasets and perform operations	HEDF Toolbox

The main role of the HEDF component is within the Use Case flows of the Data Consumer and Developer personas, as these are analyzed within D2.1.

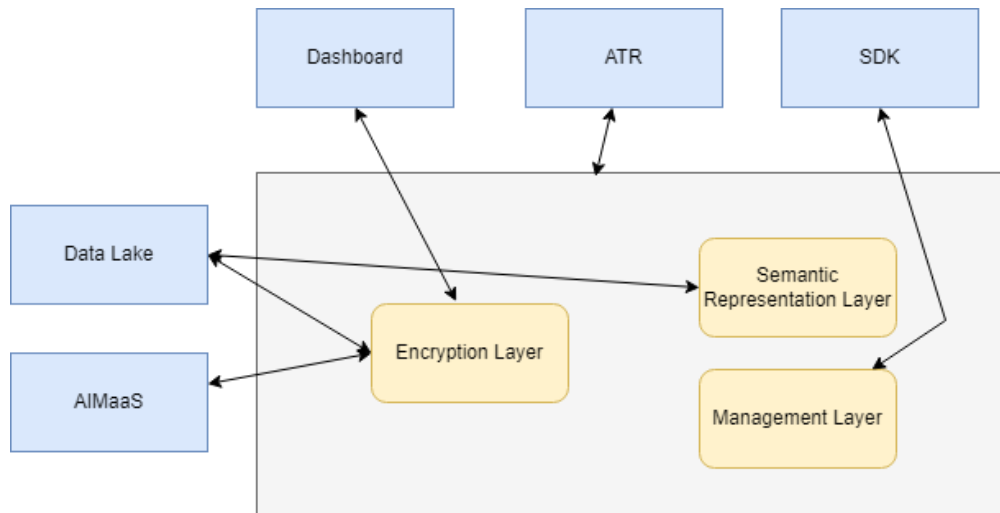


Figure 3 HEDF Architecture

Figure 3 shows the high-level architecture diagram of HEDF, along with the sub-components that comprise this ABB, and the connections and communications that are necessary across the different ABBs of the TRUSTEE ecosystem. The HEDF component consists of two main sub-components, the Encryption Layer and the HEDF middleware which consists of the Semantic Representation Layer and the Management Layer. The Encryption Layer is responsible for the encryption and fusion functionalities, as these are explained in the next sections of this report, while the middleware is mainly responsible for any indexing and semantic representation required of the datasets.

The communication and data exchange between HEDF and the other ABBs shown in Figure 3 are explained in detail within D2.1. Some of the interactions and communications of HEDF with other ABBs that are presented in D2.1 have been removed based on the ongoing work and the project's requirements specification process. Such interactions include HEDF's communication with STM, KR and SSI-HE, since it has been decided that the information exchange that was initially deemed necessary will be replaced in each case by other ABB interactions.

3.4.1. LOW LEVEL

The following section provides a description of the main modules within the HEDF ABB.

3.4.1.1. SEMANTIC REPRESENTATION LAYER

The Semantic Representation layer plays a crucial role in transforming collected data into meaningful knowledge. This layer involves the application of common and standardized semantic ontologies to the acquired data, correlating it with relevant contextual information. By utilizing ontologies, which define the relationships and meanings of concepts, the framework establishes a shared understanding of the data, enabling effective data integration and knowledge extraction.

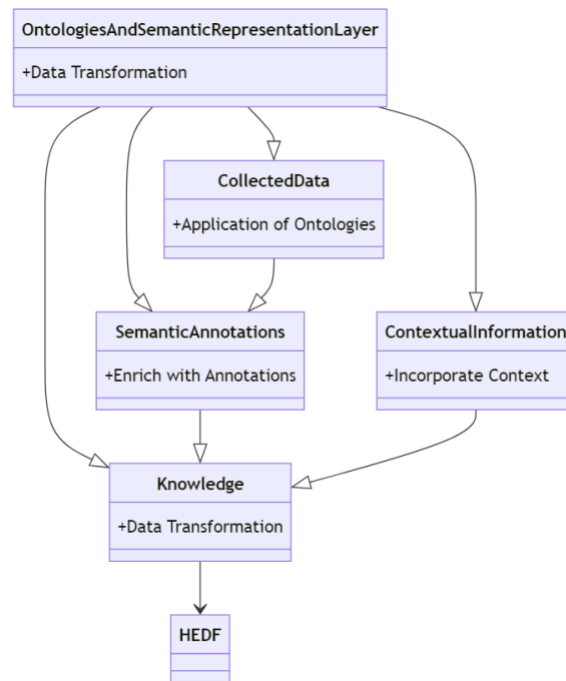


Figure 4 Semantic Representation layer

At this layer, the collected data is enriched with semantic annotations that capture the underlying semantics and relationships within the data. These annotations provide a structured representation of the data, facilitating interoperability and enabling advanced querying and reasoning capabilities. By applying semantic ontologies, the framework ensures that the data is described using a common vocabulary and follows a standardized data model, allowing for seamless integration across different domains and sources.

Furthermore, the Semantic Representation layer enables the incorporation of contextual information, which enhances the understanding of the data within its specific context. Contextual information may include location data, temporal information, domain-specific knowledge, or any other relevant metadata. By correlating the collected data with appropriate contextual information, the framework enables a deeper understanding of the data and supports more context-aware decision-making processes. By transforming collected data into knowledge through ontologies and semantic representation, the TRUSTEE framework enables advanced data analysis, inference, and reasoning. The utilization of standardized semantic ontologies ensures consistency, interoperability, and reusability of the knowledge extracted from the data. This layer serves as a bridge between raw data and actionable insights, providing a solid foundation for subsequent layers in the framework, such as data fusion and knowledge discovery.

In summary, the Semantic Representation layer leverages common and standardized semantic ontologies to transform collected data into knowledge. By applying semantic annotations and correlating data with contextual information, the framework establishes a shared understanding of the data, enhances interoperability, and enables advanced querying and reasoning. This layer serves as a critical step in extracting valuable insights from the collected data and lays the groundwork for subsequent stages in the data fusion and knowledge discovery processes.

3.4.1.2. MANAGEMENT LAYER

Future work in this layer will consist of exploring the management of the data that will derive from external data sources, e.g. GAIA-X, IDS connectors, etc. This layer will investigate the data catalogues, metadata and any other

types of information that will stem from the external connectors, and send to the other ABBs (e.g. Data Lake, SSI-HE etc.) the appropriate requested data.

The Communication Management & Abstraction layer is responsible for the setting up and provision of APIs and the investigation of the different communication protocols to allow the communication of HEDF across different ABBs within the TRUSTEE ecosystem and the communication with external data sources, such as GAIA-X, IDS etc.

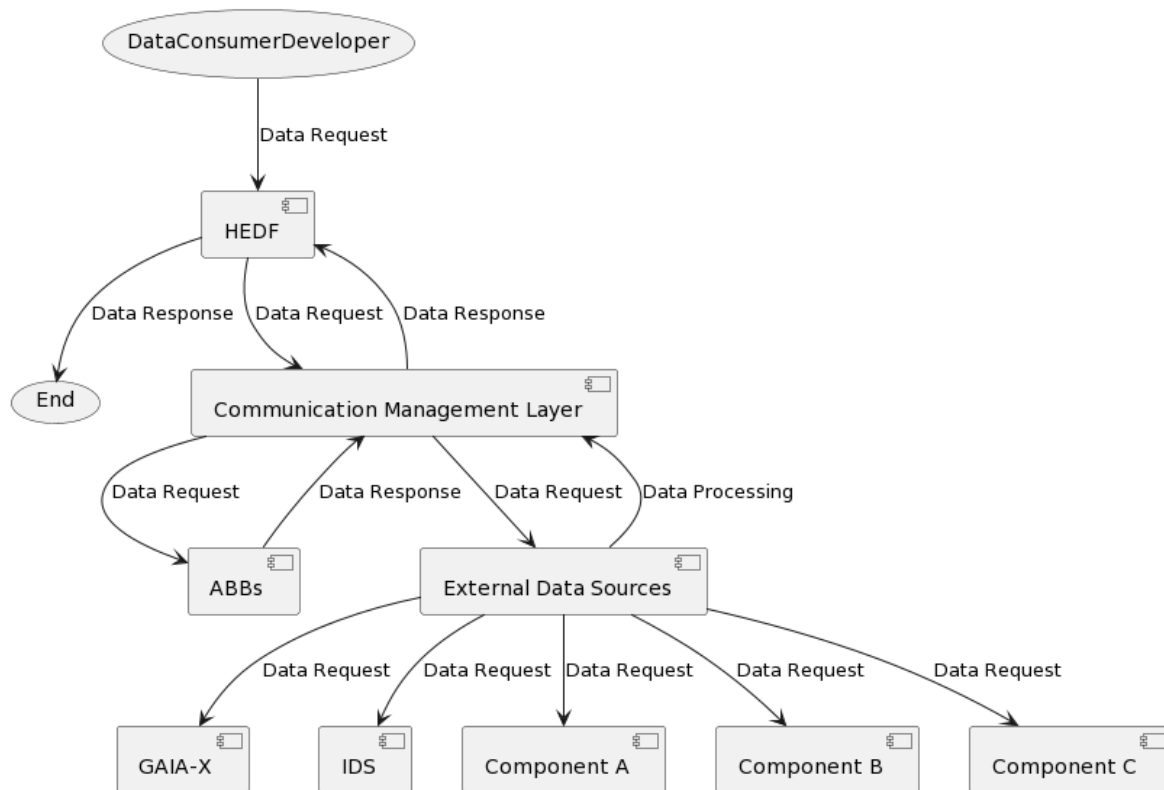


Figure 5 Low-level architecture of Communication Management Layer data flow

Figure 5 shows the data flow within the Communication Management Layer of HEDF, where data requests and responses are handled for the different ABBs and external data sources.

The following table includes the APIs that will be designed for the exchange of information, as they have been suggested so far. Future versions of the data models will be provided within D3.2.

Table 4 APIs for Communication Management within HEDF

API	Data exchanged
HEDF-Dashboard	Dashboard requests to HEDF the results of the computation that the TRUSTEE User requested
HEDF-SDK/OneSS	SDK provides to HEDF with external connectors for accessing data from external sources (ISD, EOSC, EGI connectors)

HEDF-DATALAKE	HEDF will receive from the Data Lake the stored datasets indexes / AI models
HEDF-AIMaaS	HEDF will communicate to support the HE-FL enabled processes of AIMaaS
HEDF-ATR	HEDF will send ATR records of transactions

3.4.1.3. ENCRYPTION LAYER

The encryption layer in HEDF is composed of three subcomponents as shown in figure below.

- **TRUSTEE Platform:** This subcomponent is in charge of the homomorphic keys generation (both public key for the encryption process and private key for the decryption process) as well as the suitable HE context depending on the specific algorithm to be applied. This information will be securely generated, maintaining the private key in a private and secure storage while storing the public key and the HE context on the data lake to be accessible by TRUSTEE data providers and consumers. As private key remains securely stored on TRUSTEE platform, the decryption process which requires the use of this private key also happens inside the TRUSTEE platform to maintain privacy of the private key. For this purpose, this subcomponent will contain a secure cryptography module (secure enclave) for the keys generation and decryption processes.
- **Data provider:** This subcomponent is in charge of the homomorphic encryption process with the public key and the HE context/configuration (obtained from the data lake). As a result, data provider's data set will be homomorphically encrypted in a proper way to execute the requested operation (considering the algorithm requirements as well as the data fusion need). This subcomponent will be provided as a Docker image to be run on the data providers premises taking as input the public key and the HE context/configuration.
- **Data consumer/AIMaaS user:** This subcomponent is in charge of the operations/algorithms execution on the homomorphic domain; that means, on the one or several fused homomorphically encrypted data from different data providers. As a result, the result of the operation is obtained, also encrypted, and needs to be sent back to TRUSTEE platform for the decryption process (as it is where the private key is securely stored). This subcomponent will be provided as a different Docker image for each algorithm/operation to be run on the data consumers/AIMaaS user premises taking as input encrypted data from one or several data providers.

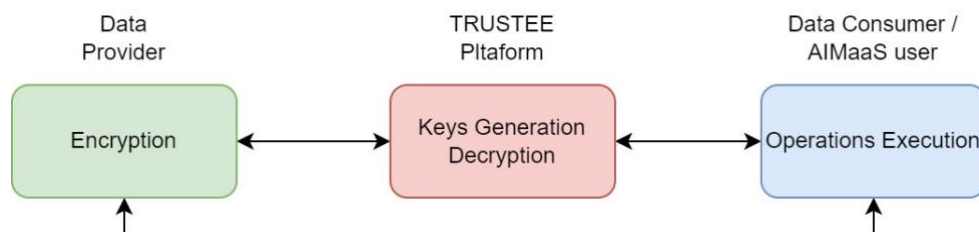


Figure 6 HEDF encryption layer architecture

Summarizing, the main functionality of this layer is to allow the execution of specific operations (defined by TRUSTEE use cases) over encrypted data, maintaining data confidentiality between data provider and data consumer as well as among data providers when different datasets from different data providers are fused.

In detail, the main functionalities of the encryption layer are:

- Secure HE keys (and HE context) generation inside a secure enclave inside TRUSTEE Platform.
- Secure keys sharing when needed (HE public key, HE context) with data providers and data consumers from TRUSTEE Platform.
- Data encryption with HE public key by data providers.
- HE encrypted data sharing from data providers to data consumers.
- Secure HE algorithms pulling from TRUSTEE platform (to be applied over encrypted data) by data consumers.
- Execution of algorithms over HE encrypted data by data consumers.
- HE encrypted result sharing with TRUSTEE platform by data consumers.
- Secure HE decryption process of the encrypted result with HE private key inside the secure enclave inside TRUSTEE platform.
- Secure HE decrypted result sharing with data consumer by TRUSTEE platform.

OPERATION FLOW

Figure 7 and Figure 8 describe the operation flow in a sequence diagram and in the architecture, respectively, of the encryption layer of HEDF considering the three subcomponents described before: TRUSTEE platform in red, data provider in green and data consumer in blue. Interactions with other TRUSTEE components have been also defined for better understanding of the complete flow.

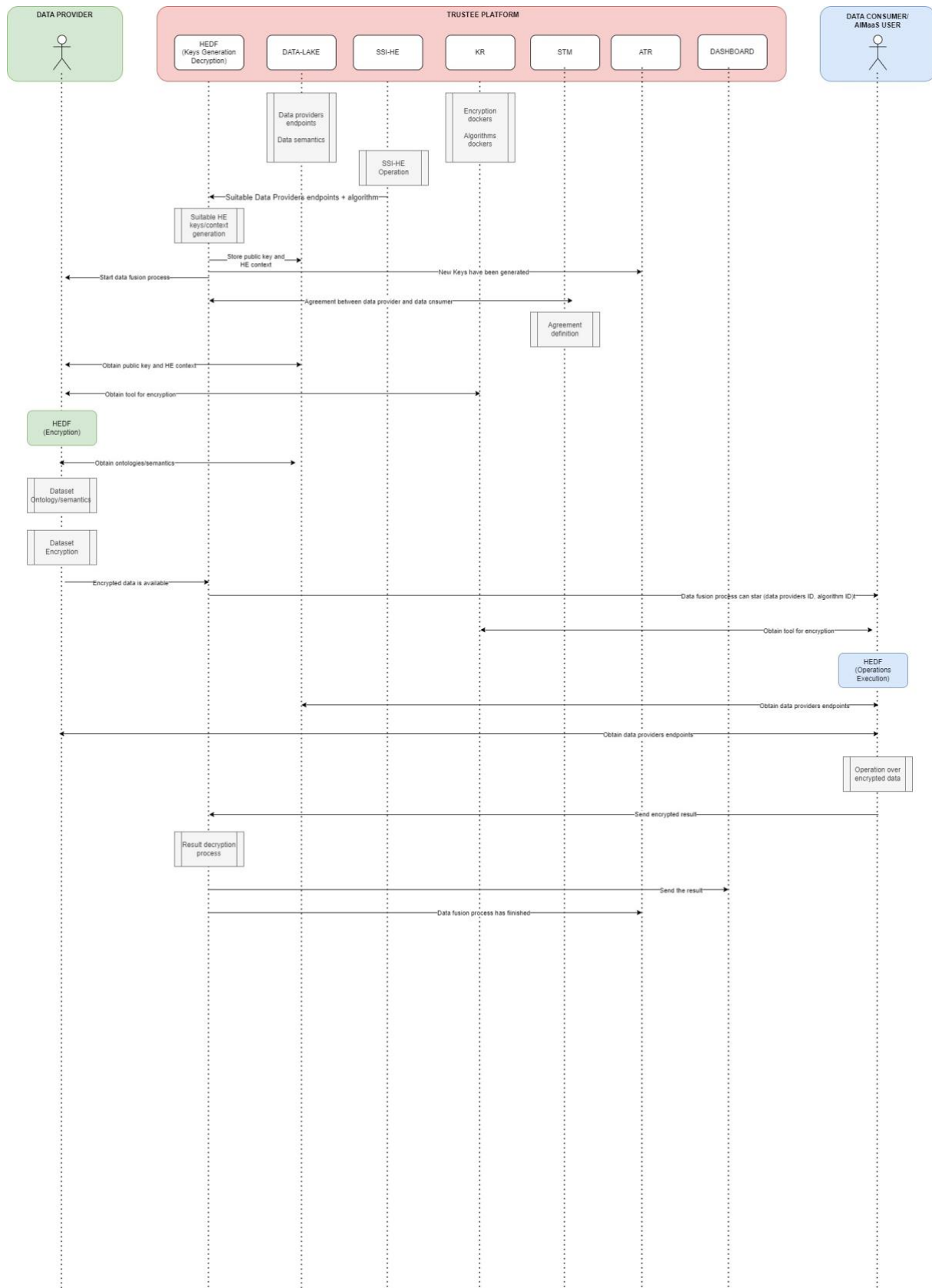


Figure 7 HEDF encryption layer operation flow in a sequence diagram

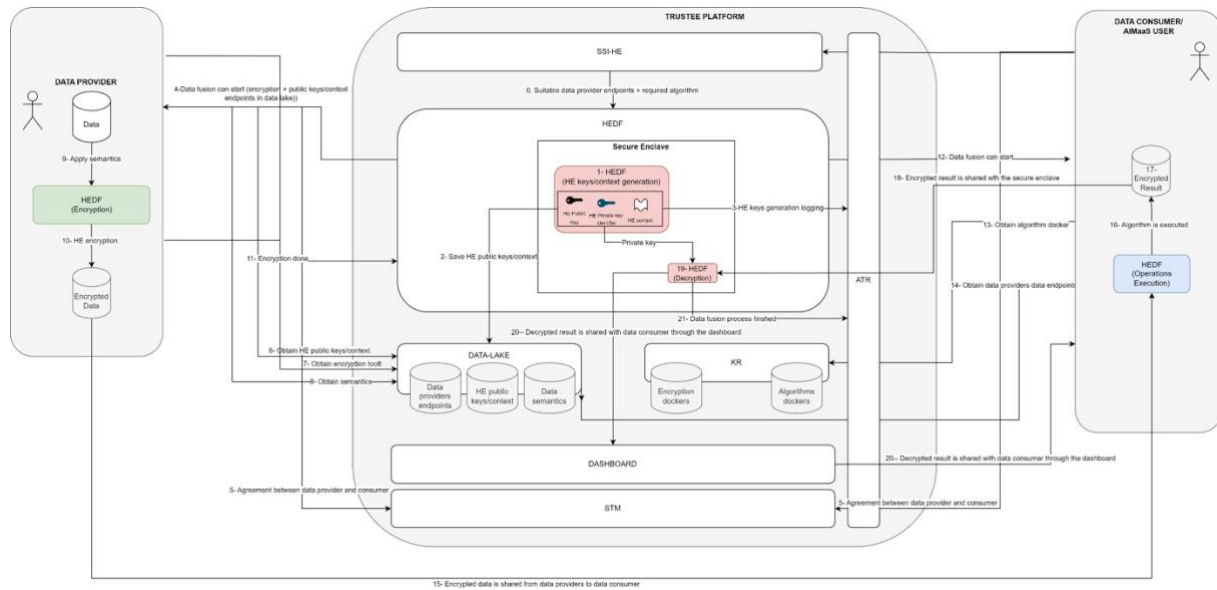


Figure 8 HEDF encryption layer operation flow in the architecture

Before starting the HEDF operation, as prerequisites, Docker images of the encryption tools (for OpenFHE and TFHE-rs) for data providers could be available on the KR. Moreover, the Docker images of the different algorithms to be applied on the homomorphic domain will be also available on the KR for data consumers.

The operation of the HEDF component starts when the SSI-HE operation finishes resulting in a list of suitable data providers endpoints and the algorithm to be applied that is shared with HEDF component. From this moment, the flow described in Figure 7 is as follows:

1. The TRUSTEE Platform HEDF subcomponent generates the necessary HE keys (OpenFHE or TFHE-rs) for the specific algorithm to be applied (private key, public key and HE context/configuration) inside the secure enclave.
2. The generated public key and the HE context/configuration are stored in the TRUSTEE data lake to be accessible by the data providers or data consumers.
3. The generation of these new HE keys is logged in the ATR.
4. HEDF informs the suitable data providers (through the endpoints received from SSI-HE) about the start of the data fusion process.
5. An agreement between data providers and data consumers needs to be defined in the STM (more details will be provided in D4.1).
6. The data provider obtains the HE public key and HE context/configuration from the TRUSTEE data lake.
7. Additionally, the data provider obtains the necessary encryption tool (Docker) from the KR.
8. Finally, the data provider obtains the semantics/ontologies to be applied to its dataset for the specific operation from the semantic layer of the HEDF component.
9. The data provider applies the ontologies and semantics (suitable format) to its own data.
10. The data provider encrypts his dataset with updated ontologies (from step 9), using the encryption tool (from step 7) and using the obtained configuration and public key (from step 6).
11. The data provider notifies TRUSTEE Platform HEDF subcomponent when data has been properly encrypted.
12. When all required data providers notifications have been received, TRUSTEE Platform HEDF subcomponent informs the data consumer about the start of the data fusion for him.
13. The data consumer obtains the Docker with the specific algorithm to be applied from the KR.

14. The data consumer obtains the data endpoints of the suitable data providers from the TRUSTEE data lake.
15. Based on data endpoints from step 14, encrypted data of the different data providers is obtained by the data consumer.
16. The specific operation is executed on the algorithm tool from step 13 over encrypted data from step 15.
17. An encrypted result is obtained by the data consumer.
18. This result is shared with the TRUSTEE Platform HEDF subcomponent for its decryption inside the secure enclave.
19. The result is securely decrypted inside the secure enclave of the TRUSTEE Platform HEDF subcomponent.
20. The decrypted result is shared with the data consumer through the dashboard.
21. The end of the data fusion process is logged in the ATR

3.5. IMPLEMENTATION

3.5.1. FUNCTIONAL DESCRIPTION

The following sub-sections include the implementation presentation of the Encryption Layer, as it has been realized so far. The other two HEDF layers will initiate their implementation phase and reported in the next versions of this live document.

3.5.1.1. ENCRYPTION LAYER

Regarding the encryption layer, the focus during this first implementation phase has been on the design and theoretical analysis of technologies that could be useful for TRUSTEE (different HE libraries, secure enclaves...). Additionally, a first isolated and standalone prototype of the encryption layer (with no communications with other TRUSTEE components) has been implemented, focusing mainly on a basic complete HE processes. Additionally, the three subcomponents have not been distributed yet; all of them have been locally deployed sharing different volumes in order to easily share the needed information (HE public key, HE context, data and result). Finally, by the moment, secure enclaves have not been considered in this first implementation, so keys generation and decryption process are not secure enough as they have been implemented in traditional untrusted context (they will be considered for the next iteration of this component).

Taking everything into consideration, Figure 9 shows the functional components of the first implementation of the HEDF encryption layer. These functionalities have been implemented for the two considered HE frameworks: OpenFHE and TFHE-rs.

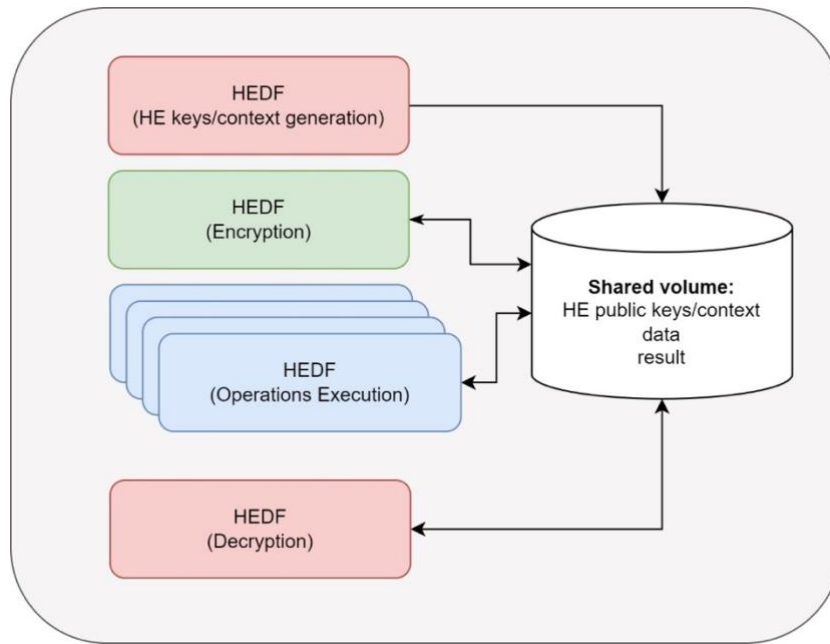


Figure 9 HEDF encryption layer first implementation functional components

In detail, the functionalities of the first implementation in comparison with the complete functionalities of the layer, defined in section 3.4.1.3 are gathered in Table 5.

Table 5 HEDF encryption layer first implementation functionalities

Functionality	Coverage Status	Current implementation
HE keys (and HE context) generation inside a secure enclave inside TRUSTEE Platform.	Partially	HE keys (and HE context) generation in a traditional untrusted context (without considering secure enclaves).
Secure keys sharing when needed (HE public key, HE context) with data providers and data consumers from TRUSTEE Platform.	No	Shared volumes are considered. No interaction with the data lake has been implemented yet.
Data encryption with HE public key by data providers.	Yes	This has been unity tested and developed
HE encrypted data sharing from data providers to data consumers.	Partially	Shared volumes are considered. Distribution has started but not completely implemented yet, and is expected to be finalized in later iterations.

Secure HE algorithms pulling from TRUSTEE platform (to be applied over encrypted data) by data consumers.	Partially	Shared volumes are considered. No interaction with the KR has been implemented yet.
Execution of algorithms over HE encrypted data by data consumers.	Partially	Algorithms for healthcare and education domains have been implemented. Table 6 for more details.
HE encrypted result sharing with TRUSTEE platform by data consumers.	Partially	Shared volumes are considered. No distribution has been implemented yet.
Secure HE decryption process of the encrypted result with HE private key inside the secure enclave inside TRUSTEE platform.	Partially	HE decryption process in a traditional untrusted context (without considering secure enclaves).
Secure HE decrypted result sharing with data consumer by TRUSTEE platform.	Partially	Shared volumes are considered. No distribution has been implemented yet.

As a summary, HE keys/context generation, HE encryption and HE decryption processes have been implemented for OpenFHE and TFHE-rs. Regarding the implemented algorithms, TRUSTEE use cases have been considered. By the moment, the health and education domains have been considered (for more details, refer to D5.1). Table 6 gathers and describes the implemented algorithms.

Table 6 HEDF encryption layer first implementation HE algorithms

Use case	Description	Comments
Health	Perform a number comparison	Implemented
Health	Perform a “lookup search” on a Homomorphic Encrypted version of “eCRF_Rapid_Covid19_Module2_2A_Vitals” to count the number of people with temperature > 39	Implemented
Health	Perform a “transitive closure” search on a Homomorphic Encrypted version of “encounter.csv” to compute the total number of people that have been infected either directly or indirectly by a given positive case (forward tracing).	Implemented (only two levels of infection have been considered due to complexity reasons).
Health	Verify that is possible to infer a linear regression on a Homomorphic Encrypted version of “eCRF_Rapid_Covid19_Module2_2A_Vitals” to establish a correlation between the temperature, the heart rate the respiratory rate, the blood pressure and the Glasgow coma score	Implemented

Education	Descriptive statistics: summarise the data to measure mean, median, mode, range, standard deviation.	Implemented
Education	Regressions: to identify any relationship between independent variables and a dependent one.	Implemented
Education	Clustering statistics – clustering algorithm to use for similar group students and compare based on their characteristics	Implemented
Health & Education	PPRL (privacy preserving record linkage) between the Health Dataset (HE version of “person.csv”) and the Education Dataset in the hypothesis that a school needs to identify a student when one of her/his parents has tested positive to COVID-19.	Implemented (this refers to KPI of objective 2)

3.5.2. OPERATIONAL DESCRIPTION

This section provides an initial overview of the operations supported by the HEDF component. Future work will be reported in D3.2. For now, the following sub-sections show a set of diagrams (e.g. sequence, data flows) that show how HEDF works to support the user scenarios of the different TRUSTEE personas.

3.5.2.1. HEDF MIDDLEWARE LAYERS

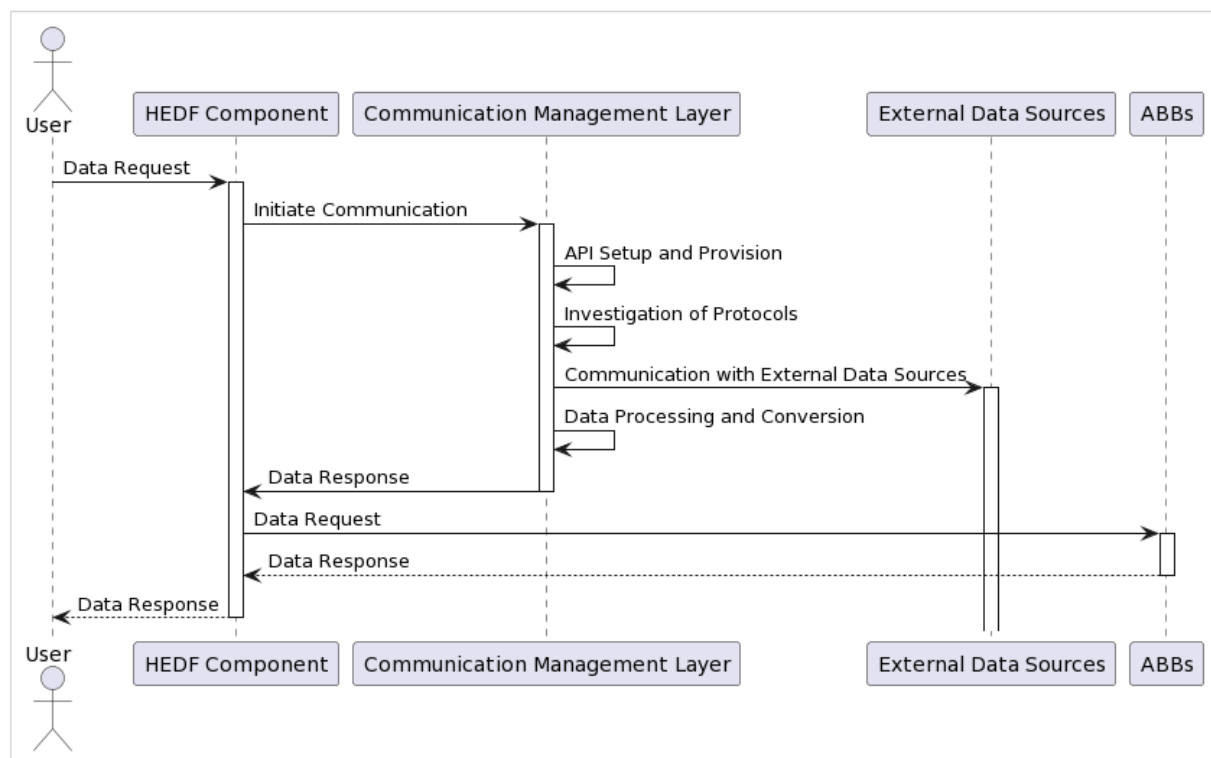


Figure 10 HEDF middleware layer sequence diagram

3.5.2.2. ENCRYPTION LAYER

Figure 11 describes the operation flow in a sequence of the first implementation of the encryption layer of HEDF, considering that all the components have been deployed in the same machine and information exchange is done through shared volumes. As mentioned, this process has been implemented for the two considered HE frameworks (OpenFHE and TFHE-rs) as well as for different algorithms (see Table 6).

By the moment, the HEDF encryption layer has been implemented as a standalone solution, so no interaction with other TRUSTEE ABBs (see section 3.4) has been implemented yet. The current operation flow is as follows: the user starts the HEDF process through the screen, and all the steps happen automatically: HE keys generation, data encryption, computation over encrypted data and data and result decryption. The result is sent back to the user through the screen.

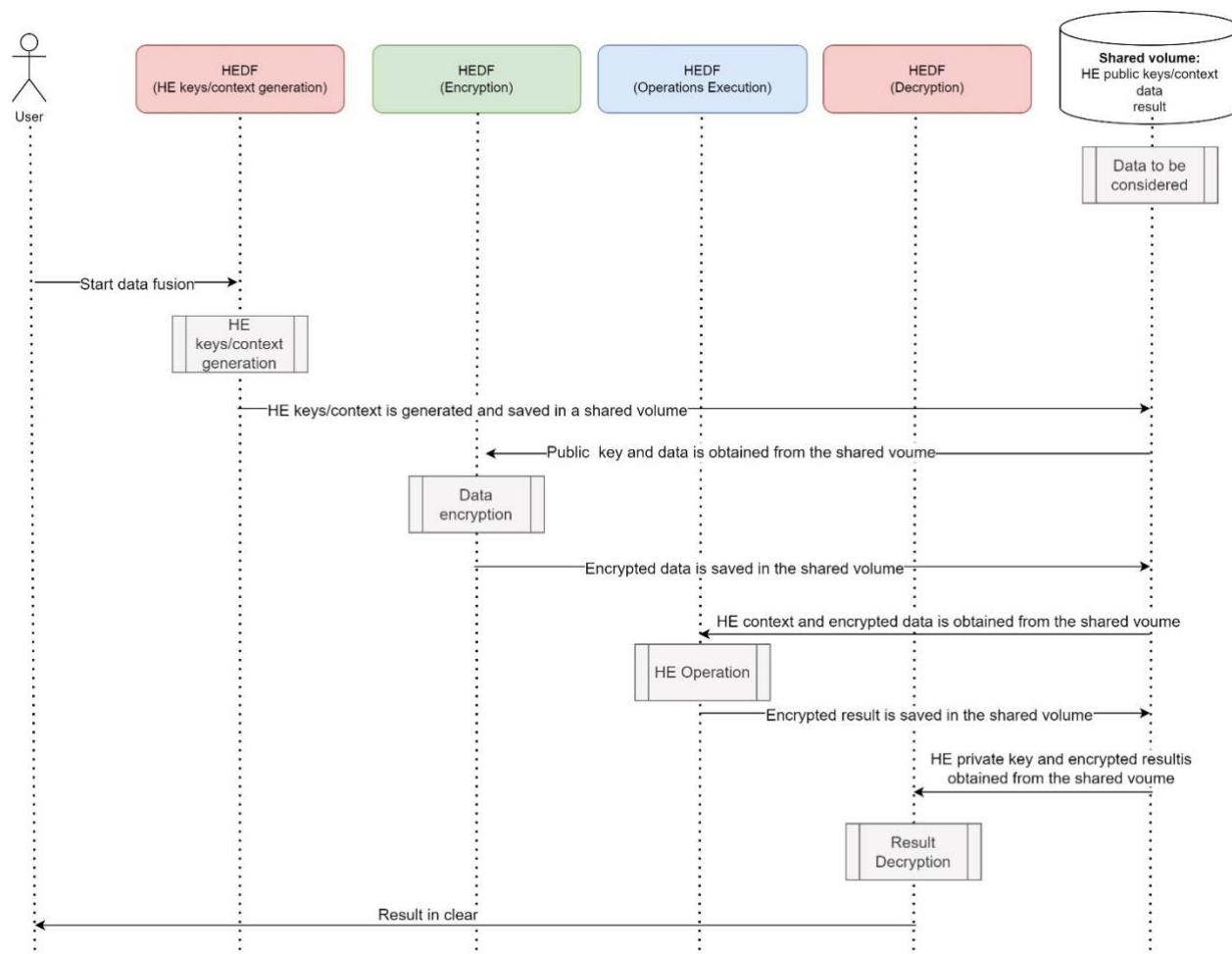


Figure 11 HEDF Encryption Layer first implementation operation flow

3.5.3. TECHNICAL DESCRIPTION

3.5.3.1. ENCRYPTION LAYER

IMPLEMENTATION DETAILS

Figure 12 shows the technical scheme of what has been implemented for the HEDF encryption layer.

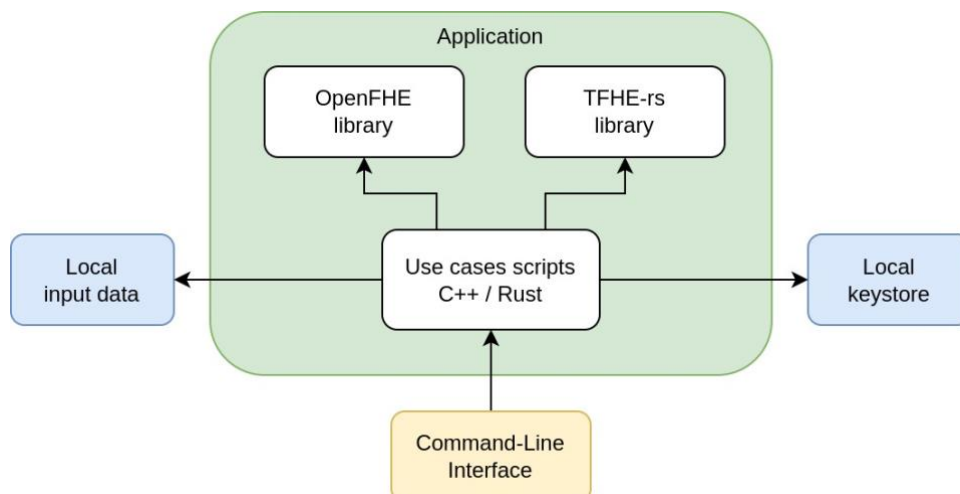


Figure 12 HEDF encryption layer first implementation details

The first implementation has been done to validate the feasibility of using homomorphic encryption in the use cases. Most of the components require integration with other TRUSTEE components still under development, which is why this preliminary implementation has been done exclusively working locally.

In any case, the components in Figure 12 are explained below to better understand how the tool works:

- **Local input data.** It is a local repository (shared volume) to temporarily store the keys and contexts needed for the executions.
- **Local keystore.** It is a local repository (shared volume) to store the clear data that is going to be encrypted and used as input data in the executions.
- **Application.** It is composed by:
 - Use cases scripts. It is a set of scripts that contains the algorithms considered in TRUSTEE. They use of the homomorphic cryptography libraries previously installed in the operating system. As explained before, for this first implementation, all functionalities (HE keys/context generation, encryption, operation and decryption) run in sequence in these scripts.
 - OpenFHE library. Framework used to perform the necessary homomorphic cryptographic operations with the CKKS and BGV schemes. It is installed on the machine running the cryptographic processes.
 - TFHE-rs library. Framework used to perform the necessary homomorphic cryptographic operations with the TFHE scheme. It is installed on the machine running the cryptographic processes.
- **Command-Line Interface.** It is an interface that allows users to interact with a system or application via commands.

Each use case script includes the implementation of a specific algorithm. The details of the algorithms already implemented in this first implementation phase are as follows:

- Number comparisons (health domain)

This test uses the TFHE-rs library to compare numeric values from a data input via CLI. It encrypts the values independently using the public key, and calculates the maximum value, the minimum value, the total sum of all of them, the total product, and, finally, it performs a greater than and equal than a specific value comparison.

- Look up search (health domain)

This test uses TFHE-rs to perform a comparison operation against a set of ciphertexts, obtaining as a result the number of elements in the set that meet the condition of being greater than 39. The input data can be of any type as long as it is an operation allowed in TFHE-rs. For the health use case, this algorithm is considered for detecting patients with excessive temperature.

- Transitive closure (health domain)

This test is also based on TFHE-rs, and uses as input a database that indicates the close contacts that each patient has had, so that the relationships between them can be correlated and grouped.

- Linear Regression (health domain)

This test is operated using OpenFHE, in particular the CKKS scheme to work with decimal values. The objective of the test is to find out if it is possible to homomorphically calculate the coefficients of a linear regression, training with a dataset with two attributes, mutually related and homomorphically encrypted. Subsequently, the obtained line coefficients are deciphered and encrypted again to apply the inference process on a different dataset with similar characteristics, which allows homomorphically predicting the results of one of the attributes from the other.

- Descriptive statistics (education domain)

In this case, the technology used in the test is OpenFHE with a CKKS scheme and a series of statistical operations are performed on a homomorphically encrypted dataset. The implemented operations are the mean, variance, standard deviation, median, and quartiles.

One limitation of operating on encrypted values is that decisions cannot be made based on the results obtained homomorphically, since they are also encrypted. This could be a problem since to perform some calculations a data preprocessing could be needed. For example, to calculate the median it is necessary to order the input data. For this reason, this algorithm requires some preprocessing before encryption.

- Logistic regression (education domain)

This is a test performed with the OpenFHE cryptography library and the CKKS scheme. The objective of the test is to find out if it is possible to homomorphically calculate the training and the inference of a logistic regression, in both cases with encrypted data inputs.

The training can be done using an operation provided by OpenFHE called EvalLogistic, which calculates a polynomial approximation of the logistic function to obtain the encrypted weights. This function has an important limitation as it is necessary for the data to be normalized in order to properly configure the EvalLogistic function, otherwise, the obtained result will not be correct.

In addition, to perform many iterations on the training data, it is necessary to greatly increase the *multdepth* parameter (number of times that can be multiplied over the same ciphertext), which determines the amount of data that can be trained with, and the security level used to configure the cryptographic context.

Once the weights are obtained, the inference algorithm can be applied.

- Clustering (education domain)

This test allows a clustering algorithm to be carried out on a set of encrypted data using the TFHE-rs library. The applied clustering algorithm is K-means, but only in the first iteration, where the number of clusters to be obtained is indicated and the distance from each entry to each of the clusters is homomorphically calculated. Then, the cluster with the smallest distance is homomorphically obtained.

- Datasets correlation (health and education domain)

This test considers as input two databases encrypted homomorphically and with the same identifiers, one of them contains private medical data of a group of students from a school, and the other one contains personal information of these students such as the ID numbers. The algorithm starts looking for students who meet the condition of having a temperature greater than 39° (look up search from health domain). However, it uses the student identifiers to compare them with the second database, and thus homomorphically obtain the ID numbers of the students who meet the condition.

3.5.4. TECHNICAL DETAILS

The Technical Details section provides an in-depth overview of the development, installation, and deployment aspects of the TRUSTEE platform. This section is crucial for technical stakeholders and development teams to understand the tools, technologies, and infrastructure required to successfully implement and deploy the HEDF ABB.

3.5.4.1. HEDF MIDDLEWARE LAYERS

Table 7 gathers some technical details about the ongoing implementation phase.

Table 7 HEDF middleware layers technical details

DEVELOPMENT	
Type (Software/Hardware)	Software
Operating Systems	Linux
Database and databases tools	Apache2 Web Server for CDN and files delivery
GUI	RESTful API endpoints
Programming language	Python
Development Environment	Linux & Windows Hybrid
INSTALLATION AND DEPLOYMENT	
Software Requirements	Open Source Software mainly
Hardware Requirements	Linux Server
Containerization	Feasible with Docker and/or VMs and other technologies.
Communications	Sockets, HTTP, REST, JSON, OAuth2 etc.

3.5.4.2. ENCRYPTION LAYER

Table 8 gathers some technical details of the first implementation phase.

Table 8 HEDF encryption layer first implementation technical details

DEVELOPMENT	
Type (Software/Hardware)	Software
Operating Systems	Linux (Ubuntu)
Database and databases tools	File System
GUI	Yes
Programming language	C++, Rust
Development Environment	Visual Studio Code
INSTALLATION AND DEPLOYMENT	
Software Requirements	Docker, OpenFHE, TFHE-rs
Hardware Requirements	At least 16GB RAM
Containerization	Docker
Communications	-

3.6. VALIDATION

This chapter includes information on the validation process for the encryption layer. The validation results for the other HEDF middleware layers will be showcased in D3.2, as the main focus of the first round of work was centred around the encryption process.

3.6.1. ENCRYPTION LAYER

Some functional tests have been defined for this first implementation phase.

- **Functional Test 1: OpenFHE keys generation.** Figure 13 shows a screenshot of the cryptographic material generated in OpenFHE and their sizes. This includes crypto context, public key and two private keys for FHE threshold functionality.

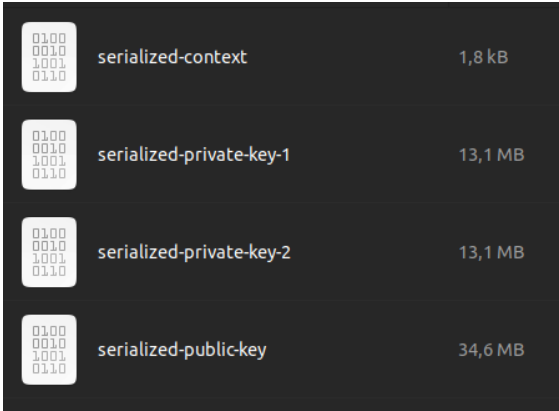


Figure 13 HEDF encryption layer first implementation functional test 1

- **Functional Test 2: TFHE-rs keys generation.** Figure 14 shows a screenshot of the keys generated in TFHE-rs and their sizes. It s important to highlight the large size of the TFHE public key.

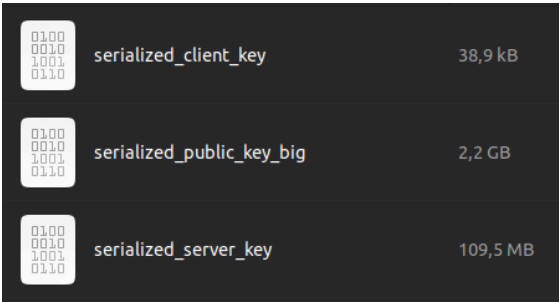


Figure 14 HEDF encryption layer first implementation functional test 2

- **Functional Test 3: Number comparisons (health domain).** Figure 15 shows a screenshot of the number comparison test from the health domain. Here, all the functionalities related to the encryption process are shown: keys generation, encryption data, executing the algorithm and result decryption.

```

#####
Running Test 0: Number comparisons
#####

  Compiling healthcare use case v0.1.0 (/home/prt1185a/Escritorio/TRUSTEE/3wp/uc_demos/healthcare)
  Finished release [optimized] target(s) in 3.09s
  Running `target/release/healthcare use case 0`
2023-07-21 11:02:56 [INFO] Generating keys...
2023-07-21 11:03:00 [DEBUG] Duration to generate client, server and public keys: 4217 ms

Please, enter a number:
5

Please, enter a number:
40

Please, enter a number:
30

Please, enter a number:
10
2023-07-21 11:03:35 [INFO] Encrypting data...
2023-07-21 11:03:39 [DEBUG] Duration to encrypt the 4 inputs: 4175 ms
2023-07-21 11:03:39 [INFO] Executing algorithm...
2023-07-21 11:04:20 [DEBUG] Duration to execute the algorithm: 40934 ms
2023-07-21 11:04:20 [INFO]
    The minimum value is => 5

2023-07-21 11:04:20 [INFO]
    The maximum value is => 40

2023-07-21 11:04:20 [INFO]
    The total aggregation is => 85

2023-07-21 11:04:20 [INFO]
    The total product is => 60000

2023-07-21 11:04:20 [INFO]      5 is not equal to 40
2023-07-21 11:04:20 [INFO]      40 is greater than 30

```

Figure 15 HEDF encryption layer first implementation functional test 3

- **Functional Test 4: Look up search (health domain).** Figure 16 shows a screenshot of all the functionalities related to the encryption process for the look up search test from the health domain.

```

#####
Running Test 1: Look up search, temperature > 39
#####

  Finished release [optimized] target(s) in 0.24s
  Running `target/release/healthcare use case 1`
2023-07-21 11:05:59 [INFO] Generating keys...
2023-07-21 11:06:02 [DEBUG] Duration to generate client, server and public keys: 3837 ms
2023-07-21 11:06:07 [INFO] Reading DB...
2023-07-21 11:06:07 [INFO] Reading file: data/health_data.csv
2023-07-21 11:06:07 [INFO] Encrypting data...
2023-07-21 11:06:19 [DEBUG] Duration to encrypt an array of 40 values: 11832 ms
2023-07-21 11:06:20 [INFO] Executing algorithm...
2023-07-21 11:06:24 [DEBUG] Duration to execute the algorithm: 4495 ms
2023-07-21 11:06:24 [INFO] Decrypting result...
2023-07-21 11:06:24 [INFO] The number of people with fever > 39 => 7

```

Figure 16 HEDF encryption layer first implementation functional test 4

- **Functional Test 5: Transitive closure (health domain).** Figure 17 shows a screenshot of all the functionalities related to the encryption process for the transitive closure test from the health domain.


```

#####
Running Test 0: Descriptive statistics
#####

/bin/sh: 1: cd: can't cd to /home/prt1185a/Escritorio/TRUSTEE/education/build
make[2]: *** [CMakeFiles/statistics.dir/build.make:111: CMakeFiles/statistics.dir/depend] Error 2
make[1]: *** [CMakeFiles/Makefile2:86: CMakeFiles/statistics.dir/all] Error 2
make: *** [Makefile:91: all] Error 2

2023-07-21 13:51:47 [INFO] ----- EVAL STATISTICS -----
2023-07-21 13:51:47 [INFO] Loading data from file: ../data/grades.csv
2023-07-21 13:51:47 [INFO] Load complete!
2023-07-21 13:51:47 [INFO] Context configuration...
2023-07-21 13:51:56 [INFO] Encrypting data...
2023-07-21 13:52:06 [INFO] Algorithm evaluation
2023-07-21 13:52:16 [INFO] Results:
2023-07-21 13:52:16 [INFO]   Mean (Maths): 84.818182
2023-07-21 13:52:16 [INFO]   Variance (English): 24.793388
2023-07-21 13:52:16 [INFO]   Standard Deviation (English): 4.979300
2023-07-21 13:52:16 [INFO]   Q1 (Science): 81.750000
2023-07-21 13:52:16 [INFO]   Q2/Median (Science): 88.000000
2023-07-21 13:52:16 [INFO]   Q3 (Science): 90.000000
2023-07-21 13:52:16 [INFO] ----- END OF PROGRAM -----

```

Figure 19 HEDF encryption layer first implementation functional test 7

- **Functional Test 8: Logistic regression (education domain).** Figure 20 shows a screenshot of all the functionalities related to the encryption process for the logistic regression test from the education domain.

```

#####
Running Test 1: Logistic regression
#####

Consolidate compiler generated dependencies of target statistics
[ 33%] Built target statistics
Consolidate compiler generated dependencies of target train
[ 66%] Built target train
Consolidate compiler generated dependencies of target inference
[100%] Built target inference
2023-07-21 13:59:02 [INFO] ----- EVAL LOGISTIC TRAINING -----
2023-07-21 13:59:02 [INFO] Loading data from file: ../data/normalized.csv
2023-07-21 13:59:02 [INFO] Load complete!
2023-07-21 13:59:02 [INFO] The data file has 22 rows and 13 columns
2023-07-21 13:59:02 [INFO] Loading data from file: ../data/results/initial_weights.csv
2023-07-21 13:59:02 [ERROR] Could not open the file ../data/results/initial_weights.csv
2023-07-21 13:59:02 [INFO] Readed bias: 0.000000
2023-07-21 13:59:02 [INFO] Context configuration...
2023-07-21 13:59:39 [INFO] Encrypting data...
2023-07-21 14:00:05 [INFO] Algorithm evaluation:
2023-07-21 14:00:05 [INFO]   Training Logistic Regression with learning rate 0.010000 and 5 iterations
2023-07-21 14:00:05 [INFO]     Iteration 0
2023-07-21 14:03:47 [INFO]     Iteration 1
2023-07-21 14:06:34 [INFO]     Iteration 2
2023-07-21 14:08:54 [INFO]     Iteration 3
2023-07-21 14:10:45 [INFO]     Iteration 4
2023-07-21 14:12:26 [INFO] WEIGHT 0: -0.009423
2023-07-21 14:12:26 [INFO] WEIGHT 1: -0.038817
2023-07-21 14:12:26 [INFO] WEIGHT 2: -0.062541
2023-07-21 14:12:26 [INFO] WEIGHT 3: -0.017893
2023-07-21 14:12:26 [INFO] WEIGHT 4: -0.022341
2023-07-21 14:12:26 [INFO] WEIGHT 5: -0.012003
2023-07-21 14:12:26 [INFO] WEIGHT 6: -0.052768
2023-07-21 14:12:26 [INFO] WEIGHT 7: -0.037197
2023-07-21 14:12:26 [INFO] WEIGHT 8: 0.057209
2023-07-21 14:12:26 [INFO] WEIGHT 9: -0.049350
2023-07-21 14:12:26 [INFO] WEIGHT 10: -0.021931
2023-07-21 14:12:26 [INFO] WEIGHT 11: 0.001616
2023-07-21 14:12:26 [INFO] WEIGHT 12: 0.003125
2023-07-21 14:12:26 [INFO] BIAS: -0.057318
2023-07-21 14:12:26 [INFO] ----- END OF PROGRAM -----

```

Figure 20 HEDF encryption layer first implementation functional test 8

- **Functional Test 9: Clustering (education domain).** Figure 21 shows a screenshot of all the functionalities related to the encryption process for the clustering test from the education domain.

```

#####
Running Test 2: Clustering
#####

Compiling education_use_case v0.1.0 (/home/prt1185a/Escritorio/TRUSTEE/3wp/uc_demos/education)
Finished release [optimized] target(s) in 3.18s
Running `target/release/education_use_case`

2023-07-21 11:53:02 [INFO] ----- EVAL CLUSTERING -----
2023-07-21 11:53:02 [INFO] Generating keys...
2023-07-21 11:53:12 [INFO] Reading DB...
2023-07-21 11:53:12 [INFO] Reading file: data/grades.csv
2023-07-21 11:53:12 [DEBUG] Cluster 0 => [3, 90]
2023-07-21 11:53:12 [DEBUG] Cluster 1 => [10, 81]
2023-07-21 11:53:12 [DEBUG] Cluster 2 => [17, 78]
2023-07-21 11:53:12 [INFO] Encrypting data...
2023-07-21 11:53:23 [INFO] Executing algorithm...
2023-07-21 11:53:25 [INFO] Decrypting result for input [85]
2023-07-21 11:53:25 [DEBUG] Min distance: [4]
2023-07-21 11:53:26 [INFO] Closest cluster: [1]

2023-07-21 11:53:28 [INFO] Decrypting result for input [92]
2023-07-21 11:53:28 [DEBUG] Min distance: [2]
2023-07-21 11:53:28 [INFO] Closest cluster: [0]

2023-07-21 11:53:30 [INFO] Decrypting result for input [78]
2023-07-21 11:53:30 [DEBUG] Min distance: [0]
2023-07-21 11:53:31 [INFO] Closest cluster: [2]

2023-07-21 11:53:33 [INFO] Decrypting result for input [90]
2023-07-21 11:53:33 [DEBUG] Min distance: [0]
2023-07-21 11:53:33 [INFO] Closest cluster: [0]

2023-07-21 11:53:35 [INFO] Decrypting result for input [80]
2023-07-21 11:53:35 [DEBUG] Min distance: [1]
2023-07-21 11:53:35 [INFO] Closest cluster: [1]

2023-07-21 11:53:37 [INFO] Decrypting result for input [83]
2023-07-21 11:53:37 [DEBUG] Min distance: [2]
2023-07-21 11:53:38 [INFO] Closest cluster: [1]

2023-07-21 11:53:40 [INFO] Decrypting result for input [85]
2023-07-21 11:53:40 [DEBUG] Min distance: [4]
2023-07-21 11:53:40 [INFO] Closest cluster: [1]

2023-07-21 11:53:42 [INFO] Decrypting result for input [90]
2023-07-21 11:53:42 [DEBUG] Min distance: [0]
2023-07-21 11:53:43 [INFO] Closest cluster: [0]

```

Figure 21 HEDF encryption layer first implementation functional test 9

- **Functional Test 10: Datasets Correlation (health and education domain).** Figure 22 shows a screenshot of all the functionalities related to the encryption process for the dataset's correlation test from the health and education domains.

```

##### encounter.csv
Enc ID, Contact Type, Positive Case, contact
1,2,3,5
2,2,7,5
#####

Finished release [optimized] target(s) in 0.03s
Running `target/release/healthcare_use_case 4`

2023-07-21 11:17:30 [INFO] Generating keys...
2023-07-21 11:17:34 [DEBUG] Duration to generate client, server and public keys: 3851 ms
2023-07-21 11:17:39 [INFO] Reading DB...
2023-07-21 11:17:39 [INFO] Reading file: data/health_data.csv
2023-07-21 11:17:39 [INFO] Reading file: data/student_list.csv
2023-07-21 11:17:39 [INFO] Encrypting data...
2023-07-21 11:17:51 [DEBUG] Duration to encrypt an array of 40 values: 11902 ms
2023-07-21 11:18:03 [DEBUG] Duration to encrypt an array of 40 values: 11798 ms
2023-07-21 11:18:15 [DEBUG] Duration to encrypt an array of 40 values: 12218 ms
2023-07-21 11:20:24 [DEBUG] Duration to encrypt an string array of 40 values: 129094 ms
2023-07-21 11:20:24 [INFO] Executing algorithm...
2023-07-21 11:21:05 [INFO] positive case => 50345678
2023-07-21 11:21:19 [INFO] positive case => 50123456
2023-07-21 11:21:22 [INFO] positive case => 50567890
2023-07-21 11:21:30 [INFO] positive case => 50678901
2023-07-21 11:21:57 [INFO] positive case => 50890123
2023-07-21 11:22:28 [INFO] positive case => 50789012
2023-07-21 11:22:28 [INFO] positive case => 50789012

2023-07-21 11:22:41 [DEBUG] Duration to time to execute algorithm: 136374 ms

```

Figure 22 HEDF encryption layer first implementation functional test 10

3.7. NEXT STEPS

The next version of the Encryption layer of the HEDF will include:

- Implementation of the algorithms for other TRUSTEE domains (Automotive, energy and multidisciplinary)
- Distribution of the HEDF functionalities: Key generation and decryption (TRUSTEE platform), Encryption (Data provider) and Execution (data consumer) in different instances to be deployed in different machines in order to provide more realism to the solution.
- Interactions with other TRUSTEE components. The required interactions with other TRUSTEE components will be implemented.
- Continue with the validation of the algorithms in the different domains.
- Performance improvement in terms of computations.

For the rest of the HEDF layers, the next steps include:

- Finalize the data indexing framework.
- Finalize the dynamic semantic representation of new datasets arriving in TRUSTEE.
- Test and validate operations and functionalities and test communication with external data sources.

3.8. CONCLUSIONS

This chapter provided an overview of the HEDF component and its layers. This ABB consists of 3 main sub-components (encryption layer, management layer, semantic representation layer). As explained in the previous sub-chapters, the focus during this first implementation phase of the encryption layer has been on the design and theoretical analysis of technologies that could be useful for TRUSTEE (different HE libraries, secure enclaves...). Additionally, a first isolated and standalone prototype of the encryption layer (with no communications with other TRUSTEE components) has been implemented, focusing mainly on a basic complete HE processes. Interaction with other TRUSTEE components has been investigated; however, it is still a work in progress and further details will be provided in upcoming versions of this live document, namely D3.2 and D3.3.

4. PART II: AI MODEL AS A SERVICE AIMAAS

The concept of AI models as a service (AImaaS) represents a cutting-edge development in the world of artificial intelligence and data processing. This task in WP3, led by FORTH, including PASEU, K3Y, ENT, RINA, and ATHENA, is poised to revolutionize the way AI models are integrated into various applications. At its core, AImaaS aims to facilitate a plug-and-play approach within the TRUSTEE ecosystem, prioritizing seamless integration and customization of AI models to enhance data processing capabilities. One of the primary objectives of this task within TRUSTEE is to streamline the onboarding process for a wide range of AI models. This will empower TRUSTEE to adapt and extend its existing AI infrastructure hosted on the cloud. The goal is to create an environment where AI models can be effortlessly integrated, configured, and put to work to meet the specific needs of different data processing tasks. Moreover, the project will explore the feasibility of deploying AI models at the edge, catering to scenarios where critical tasks demand rapid processing that cannot be achieved solely in the cloud. The ability to leverage edge computing for specialized tasks will significantly enhance TRUSTEE's capabilities and responsiveness.

The versatility of this initiative becomes apparent when considering the diverse requirements of data processing within TRUSTEE. For instance, federated learning, which allows AI models to collaborate and analyze data locally, can be employed for fast analytics on small data sets at the source. Simultaneously, deep neural networks (DNNs) can be harnessed for complex, multi-disciplinary requests spanning multiple sources. This flexibility in AI model selection and deployment positions TRUSTEE to address a wide spectrum of data processing challenges efficiently. The overarching aim of enabling plug-and-play integration of AI models into the TRUSTEE ecosystem is to maximize the utilization of data sources. When AI models seamlessly integrate with data queries upon retrieval, the potential for extracting valuable insights and optimizing decision-making processes is greatly enhanced. This initiative's impact extends beyond the immediate realm of TRUSTEE, as it sets the stage for a paradigm shift in how organizations approach AI and data processing.

The integration of AI models as a service promises numerous benefits across various dimensions:

- **Enhanced Adaptability:** TRUSTEE's ability to accommodate a variety of AI models ensures that it can quickly adapt to evolving data processing requirements. As new models become available or existing ones are updated, they can be seamlessly incorporated into the ecosystem.
- **Customization:** The plug-and-play approach allows for tailoring AI models to specific tasks or datasets. This level of customization ensures that AI solutions are finely tuned to deliver optimal results, whether it's for image recognition, natural language processing, or predictive analytics.
- **Efficient Resource Utilization:** By deploying AI models at the edge when necessary, TRUSTEE optimizes resource utilization. Critical tasks can be processed locally, reducing latency and ensuring real-time responsiveness.
- **Scalability:** As data volumes grow, TRUSTEE's ability to run federated learning for small analytics tasks and complex DNNs for large multi-disciplinary requests ensures scalability. The system can seamlessly adapt to the demands of an expanding dataset.
- **Data Source Maximization:** The integration of AI models into data queries upon retrieval ensures that data sources are utilized to their fullest potential. This not only enhances the quality of insights but also promotes efficient data management practices.
- **Cost Efficiency:** With the ability to select and configure AI models based on specific requirements, TRUSTEE can avoid unnecessary resource allocation, leading to cost savings in cloud computing resources.
- **Competitive:** By staying at the forefront of AI integration and data processing capabilities, TRUSTEE gains a competitive edge in industries where data-driven insights are paramount.

Our efforts to establish AI models as a service within the TRUSTEE ecosystem represents a strategic move towards achieving greater efficiency, adaptability, and competitiveness. The ability to seamlessly onboard a variety of AI models, tailor them to specific tasks, and even leverage edge computing for critical operations positions TRUSTEE as a leader in the field of AI-driven data processing. This forward-thinking approach not only benefits TRUSTEE but also has the potential to transform how organizations harness the power of AI to extract insights and drive innovation. As the AIaaS project continues to evolve, its impact on data processing and decision-making will undoubtedly become increasingly profound.

4.1. INTRODUCTION

4.1.1. NEED AND OBJECTIVES

Artificial Intelligence (AI) as a Service is a rapidly evolving field that aims in transforming on how business and technology offering will be offered in the future. As AI technologies continue to advance, their integration into cloud-based services is becoming more prevalent. This section will provide an in-depth exploration of how current trends, needs, and objectives of AI as a Service are including, studied and implemented within TRUSTEE.

AI As a service

To define AI as a Service, often abbreviated as AIaaS, we need to refer to the ability of the delivery of artificial intelligence capabilities and resources through internet based (i.e., cloud) computing platforms and ecosystems. The goal is to enable organizations and developers to access AI technologies, including machine learning models, natural language processing, computer vision, and more, without the need to build and maintain their own AI infrastructure.

AIaaS is significant for several reasons and below we provide such a list.

- Make AI accessible to a broader range of organizations, from startups to large enterprises, by reducing the barriers to entry.
- AIaaS can scale resources up or down as needed, enabling businesses to handle varying workloads and demands.
- Instead of investing in AI infrastructure and expertise, organizations can pay for AI services on a subscription or usage-based model.
- AIaaS fosters innovation by allowing developers to experiment with AI technologies without the overhead of managing infrastructure.
- AI ethics and bias become increasingly important, AIaaS providers often incorporate tools and features to address these concerns.

The need and objectives for AI as a Service (AIaaS) arises from several critical factors and challenges in the adoption and utilization of artificial intelligence (AI) technologies. Below we provide a short list and explain AIaaS addresses the need for accessible, cost-effective, scalable, and customizable AI solutions. It enables organizations to harness the power of AI without the complexity and financial barriers associated with building and maintaining AI infrastructure in-house. AIaaS plays a crucial role in democratizing AI and driving innovation across various industries. These needs include but not limited to:

- a) Accessibility to AI sources in order to bridge this gap by providing readily available AI capabilities, making AI accessible to a broader audience. AIaaS aims in providing a uniform and robust way to access such services
- b) Green and cost-efficiency of building and maintaining AI infrastructure can be expensive adapting the way these services can be offered in a cost efficient and power consumption efficient way. AIaaS platforms allow organizations to avoid substantial upfront costs associated with AI hardware and software

c) Scalability and elasticity to be able to compensate with the demand for AI resources that may vary significantly over time. AIaaS platforms offer the flexibility to scale AI resources up or down based on an organization's needs to ensure that organizations can efficiently handle changing workloads without the burden of managing hardware.

d) Rapid Development and resource optimization to develop and deploy AI solutions quickly. AIaaS provides pre-built (pre-trained) AI models and tools that accelerate the development process, allowing organizations to stay competitive by bringing AI-powered applications to market faster. AIaaS is also able to allow organizations to optimize resource utilization by only using and paying for AI resources when needed.

e) Customization of solutions to address specific needs and challenges without disrupting normal operation (i.e. in Healthcare). AIaaS platforms must offer the flexibility for organizations to fine-tune AI models and algorithms to align with their unique requirements. AIaaS is able to abstract the technical complexities of AI infrastructure management, including installation, maintenance, and updates, to service providers.

f) Security Risk Mitigation: AIaaS providers typically implement robust security measures and adhere to data protection regulations, helping organizations mitigate risks related to data privacy and security.

g) Democratization of AI and Ethical AI services, ensuring that is not confined to a select group of experts or large organizations. It empowers a broader range of users, including business analysts and developers, to create and deploy AI solutions, fostering innovation and creativity. AIaaS providers are increasingly incorporating tools and features to detect and mitigate biases, promoting the development and deployment of more ethical AI systems.

h) Drive Innovation and Enhance Decision-Making per and across sectors. AIaaS encourages innovation by enabling experimentation with AI capabilities without the burden of infrastructure management, fostering new use cases and applications. By providing AI-powered insights and predictions, AIaaS empowers organizations to make data-driven decisions that can drive efficiencies and competitive advantage.

Federated Learning

Federated learning (FL) is a collaborative training paradigm for cooperatively learning a global model without sacrificing data privacy. Training is achieved by only exchanging and aggregating model parameters/gradients while training data are kept private. TRUSTEE, via the AIMaaS/FL module will provide an additional layer of security by employing homomorphic encryption in the FL aggregation step, to eliminate the potential of parameters/gradients being utilized for inferring the (private) training data of the FL participants. This trustworthy functionality will offer TRUSTEE users a secure platform for training cutting-edge AI models by sharing collaboratively obtained knowledge in a privacy preserving fashion. To achieve its goals, the AIMaaS/FL module will a) define a suitable set of attributes enabling TRUSTEE FL Prosumers to identify compatible users to participate in an FL training scheme; b) implement several HE-enabled FL aggregation operations, including traditional and personalized flavours of FL, and investigating extended fusion rules in synergy with the xAI-by-design functionality of the AIMaaS/Explainable AI module; and c) keep track of the training mechanism and the produced global models in the course of multiple FL rounds, managing also communications with the UI and xAI modules of AIaaS, as well as other ABBs, including HEDF and DL

Explainable AI

TRUSTEE can incorporate pre-trained AI models from various developers through AIMaaS. However, this kind of incorporation requires the necessary trust and explanations in a production environment. Explainable Artificial Intelligence (XAI) serves to meet this need in the field of AI and Machine Learning. The primary needs of XAI include fostering transparency, trust and accountability in AI systems. In pursuit of these needs, XAI's main objectives are to enhance transparency, improve user understanding, build trust, enable accountability, detect and mitigate bias, and enhance model reliability. Through these objectives, XAI seeks to make AI systems more

interpretable, accountable, and beneficial to individuals and society at large. The application of XAI techniques provides insights into the model's decision-making process. AI models often exhibit complexity, and understanding the rationale behind a model's specific decision, particularly when employing opaque black-box algorithms, can pose a significant challenge for humans. XAI attempts to improve transparency by providing human-understandable information in the form of visualizations or natural language explanations (Weller 2017). A core objective of XAI is to enhance trustworthiness. Trust plays an important role in fostering the broad acceptance and utilization of AI, particularly in critical domains where AI decisions can wield substantial influence (Dosilovi 2018). When the involved stakeholders can gain insight into the rationale behind AI model decisions, it significantly increases the likelihood of their acceptance and utilization of these models. As AI becomes increasingly integrated into social environments, the pursuit of explainability in these systems aims to establish fairness. In some instances, AI models have been trained on biased data, potentially leading to discriminatory outcomes (Kim 2018). Through the comprehension of the way that the model leverages input data and revealing the relationships that drive specific outcomes, XAI can highlight any instances of biased data utilization and unfair decision-making. Thus, XAI strives to mitigate the unethical and unjust deployment of AI models. Another key objective of XAI is to ensure accountability through the ability to trace how the system reaches a specific decision. Lastly, XAI plays a pivotal role in ensuring the reliability of AI models, ensuring their consistent and accurate performance over time (Khan 2022).

The application of explainability techniques in various disciplines has been significantly increasing in recent years. The detailed literature review presented in D2.1 introduced a summary of how XAI techniques are used in critical domains like Energy, Automotive, Health, Education, and Space. In the energy sector, XAI helps to identify the crucial features influencing power generation forecasting and building energy consumption, enabling more efficient energy management. In the automotive industry, it enhances transparency in computer vision models, particularly vital for self-driving vehicles, by explaining how these models interpret road and traffic sign data, ultimately promoting safety and trust. In healthcare, XAI fosters trust in AI-based diagnostic tools by providing clear explanations for decision-making, offering insights into why specific diagnoses were reached. In education, XAI ensures compliance with fairness and ethics regulations in AI-driven learning platforms, preserving integrity in personalized education. Additionally, in space exploration, XAI improves the interpretability of AI models used in satellite data analysis and spacecraft operations, enhancing mission success by shedding light on complex data-driven insights. In conclusion, explainable AI applications hold significant promise in these domains. Nevertheless, there is a clear imperative for continued research and development to fully unlock and harness their potential for practical implementation.

4.1.2. INNOVATION

4.1.2.1. AI AS A SERVICE

AI Integration in Cloud Platforms

Major cloud providers, including Amazon Web Services (AWS), Microsoft Azure, and Google Cloud, are continually enhancing their AI offerings. These platforms provide a wide array of AI services, from pre-trained models to infrastructure for training custom models.

Industry-Specific AI Solutions

AI as a Service is becoming more specialized for various industries. For instance, healthcare providers can access AI solutions tailored for medical imaging analysis, while financial institutions can leverage AI for fraud detection and risk assessment.

Democratization of AI

AI is no longer the exclusive domain of data scientists. AI platforms are increasingly user-friendly, allowing business analysts and developers with limited AI expertise to create and deploy AI solutions.

Edge AI

Edge computing, where AI processing occurs closer to data sources, is gaining traction. AI as a Service is extending to the edge, enabling real-time decision-making in applications like autonomous vehicles and IoT devices.

Explainable AI (XAI)

The need for transparency and interpretability in AI is driving the development of Explainable AI solutions. AiaaS providers are incorporating tools that help users understand AI model decisions.

AI Ethics and Bias Mitigation

Addressing ethical concerns and biases in AI models is crucial. AiaaS providers are implementing features to detect and mitigate biases and ensure AI systems align with ethical principles.

AI-optimized Hardware

Custom hardware, like GPUs and TPUs (Tensor Processing Units), is essential for AI acceleration. Cloud providers are offering specialized AI-optimized hardware as part of their AI services.

AI Marketplaces

AI marketplaces are emerging, connecting AI service providers with businesses seeking AI solutions. These marketplaces offer a wide range of AI capabilities, from chatbots to computer vision APIs.

4.1.2.1. FEDERATED LEARNING

Via the Homomorphically Encrypted Federated Learning (HE-FL) paradigm of TRUSTEE, diverse users (SMEs, industry, Universities, RTOs, policy makers, and the society at large), acting as Model Consumers/FL prosumers, collaborate in a secure and privacy-preserving fashion to enhance the performance of their AI models, without exchanging data. Utilizing this trustworthy functionality of TRUSTEE, the involved stakeholders will have access to a secure platform for producing reliable and highly performing AI models (by leveraging knowledge from otherwise inaccessible data), to improve research, services, products, and decision making in multiple domains (e.g., automotive, space, energy).

4.1.2.2. EXPLAINABLE AI

The XAI subcomponent introduces a novel capability by providing TRUSTEE personas with additional insights into the explanations behind the results of AI models. This capability enhances their ability to understand, interpret, and trust the AI-driven results generated within the platform, fostering a deeper and more informed engagement with AI technologies. In particular, the innovation aspects of the XAI subcomponent within TRUSTEE are about (a) improving trust of pre-trained AI models, (b) enhancing decision making, (c) regulatory compliance and (d) error detection. More specifically, the XAI subcomponent can enhance the trust of the TRUSTEE personas about the outcomes and operation of the AI models, providing clear explanations. Next, XAI can assist the TRUSTEE personas in decision-making, providing interpretable insights about the decisions of the AI models. Finally, in industrial environments with strict regulations, such as a healthcare environment, the XAI subcomponent can support the organisations to comply with regulations that require transparency and accountability in AI-driven

decision-making processes. Finally, the XAI subcomponent can play a significant role in the identification of errors and biases in the decisions of the AI models, allowing the TRUSTEE personas to rectify these issues before resulting in significant consequences or legal challenges.

4.2. BACKGROUND

4.2.1.1. AI AS A SERVICE

Artificial Intelligence (AI) stands as one of the most actively discussed technologies, holding the potential to significantly contribute to individual well-being, organizational success and innovation, and societal prosperity and progress⁴. According to the McKinsey Global Institute, the adoption of AI could lead to an additional global economic output of USD 13 trillion by 2030 (Thiebes 2020). Organizations are increasingly turning to AI to handle complex tasks that were traditionally considered the exclusive domain of humans. In certain the analysis of medical data to aid physicians in making faster and more accurate treatment decisions (Madani 2018) and the rapid analysis of large volumes of video footage to support criminal investigations (Crawford 2019). However, a significant challenge organizations face is the intricate and resource-intensive process of adopting and integrating AI (Dutta 2018). This hesitancy is primarily due to the scarcity of AI experts (Chui 2018), the inability of many organizations to establish and maintain the IT infrastructure required (Romero 2019), and a limited understanding of how to effectively deploy and configure AI-based systems (Wang 2018). Consequently, a large number of organizations continue to struggle with adopting AI and fully capitalizing on its potential. To facilitate the wider adoption and application of AI, cloud providers such as Amazon, Google, IBM, Microsoft, Salesforce, and SAP have begun offering machine learning, deep learning, analytics, and inference as a service, effectively translating the discourse on provisioning AI capabilities from the cloud into practical solutions. Furthermore, small and medium-sized enterprises (SMEs) are following suit, offering specialized cloud-based AI services tailored to the unique requirements of SMEs in various industries.

In essence, AIaaS combines AI, defined as the ability of a machine to perform cognitive functions associated with human minds, with the cloud computing model known for its capacity to provide ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned (Rai 2019). The primary objective of AIaaS is to make AI accessible and affordable to organizations of all sizes, regardless of their size, technological sophistication, or budgetary constraints. AIaaS guides users through the process of developing, deploying, or utilizing data analytics models without requiring them to master intricate algorithms or technologies (Elshaw 2018). This approach allows users to concentrate on tasks like training and configuring their AI models, enabling them to focus on their core competencies without the burden of handling installation, maintenance, and related management complexities.

Within TRUSTEE we have identified that AI model as a Service aim in orchestrating various module and components of the platform to achieve the aforementioned goal and **utilize and simplify the use of AI model as A service through TRUSTEE** (see figure below).

- ✓ Support in organising trust and building interoperable AI data spaces
- ✓ Reference for developing AI data spaces
- ✓ Privacy-preserving AI Models Inference

⁴ <https://link.springer.com/article/10.1007/s12599-021-00708-w#ref-CR47>

- ✓ Providing rich set of features to support challenges and requirements of AI in TRUSTEE
- ✓ Prescriptive guidelines for ensuring trust and interoperability between individual AI data spaces
- ✓ **AI models** for “**trusted**” data processing
- ✓ Accelerate AI/ML tasks at every stage
- ✓ Easy to use workflows/contracts for data processing

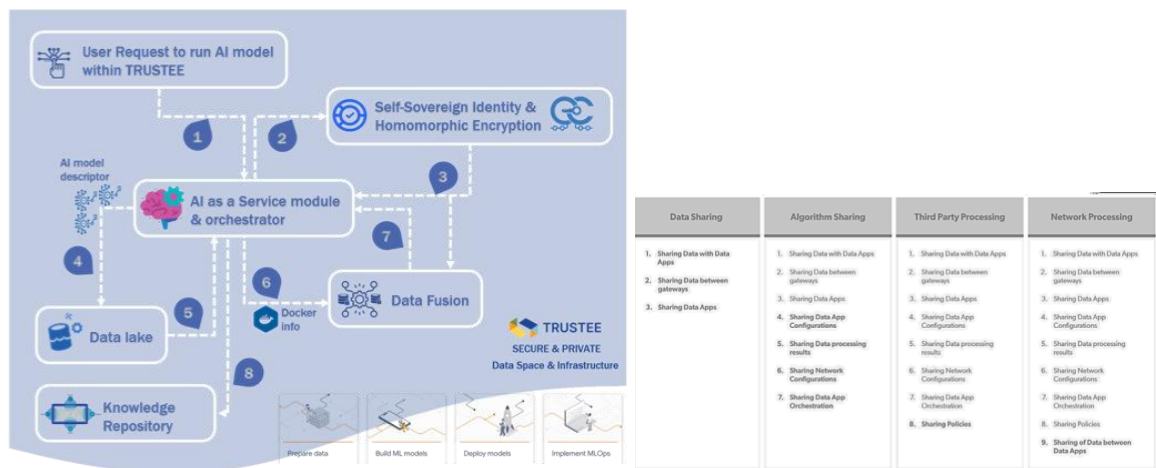


Figure 23: TRUSTEE AIaaS general component functionality and purpose

4.2.1.2. FEDERATED LEARNING

In recent years, federated learning (FL) has emerged as a new ML paradigm for achieving collaborative training and data privacy at the same time, by decoupling the provision of data at the end-user from the ML model aggregation, at the centralized server (Konečný 2016). The purpose of FL is to cooperatively learn a global model without sacrificing data privacy directly. To this end, training is performed on multiple local datasets stored at isolated data sources, without the need for collecting and processing the training data at a centralized data server. After training with local data, the participants are only required to share gradients for model updates. Then the server aggregates the gradients and transmits the updated (global) model back to each user, with the process being repeated until convergence. Thus, in FL the (global) ML model is obtained by aggregating multiple local models, each trained on its own private dataset, meaning that aggregation occurs at the model level and not at the data level, as is the case with conventional centralized ML frameworks.

This training paradigm endows FL with distinct privacy advantages compared to conventional approaches, since the personal data is stored and processed locally while only training-related parameters are exchanged. Furthermore, since the participant ID is not required by the aggregation process, updates can be transmitted without identifying metadata, utilizing de-identification methods like anonymization (Machanavajjhala 2007), obfuscation methods like differential privacy (Dwork 2006), cryptographic techniques like homomorphic encryption (Papernot 2016), and secure multi-party computation (SMC) (Yao 1986). More information on the privacy-preserving FL paradigm can be found in Part I of D2.1.

4.2.1.3. EXPLAINABLE AI

XAI refers to the set of techniques, methods, and models designed to make the decision-making processes of AI and machine learning systems more transparent, interpretable, and understandable to humans. It aims to provide insights into how AI models arrive at specific outcomes, making it possible for users to comprehend and trust these models' behaviour, detect biases, and ensure ethical and accountable AI applications.

Four main principles of XAI have been identified (Phillips 2021): Explanation, Meaningful, Explanation accuracy and Knowledge Limits. The primary principle, Explanation, emphasizes that XAI methods should provide explanations for the decisions made by AI algorithms. This principle primarily concerns the provision of explanations and doesn't inherently address the quality or accuracy of these explanations. The Meaningful principle states that XAI systems should provide explanations that are comprehensible to their users. The measurement of meaningfulness in an explanation depends on the intended audience and the specific purpose of the explanation and is an ongoing area of investigation. The Explanation Accuracy principle states that explanations must accurately represent the underlying model's decision-making process. Additionally, this principle takes into account the level of detail that explanations should encompass, a factor that varies depending again on the intended audience and the explanation's purpose. Lastly, the Knowledge Limits principle underscores that XAI should identify situations that lie beyond the scope of the AI model and where its responses may not be reliable. This principle serves the crucial purpose of bolstering trust in XAI methods by mitigating the provision of uncertain or unreliable outputs.

AI algorithms can be broadly categorized as either transparent or opaque, commonly referred to as "black-box". Transparent models allow for a comprehensible understanding of their results and the underlying process by analysing their internal structures. Nevertheless, in practical applications, the majority of AI models are not inherently transparent. This may be due to their architectures not being explicitly defined and publicly accessible, or because of their inherent complexity. Even AI experts often regard these models as black-box systems, as they struggle to gain a complete understanding of their decision-making processes. In such instances, post-hoc XAI techniques can be employed to elucidate the functioning of models that lack inherent transparency. Some common types of transparent ML models are Linear/logistic regression, General Additive Models, Rule-based learning, Decision Trees, Naive Bayes and k-nearest neighbours (KNN) while examples of post-hoc explainability methods are LIME, DeepLIFT and SHAP which are further described with more details in the section related to the technical details of the xAI subcomponent.

4.3. REQUIREMENTS

In this chapter, the compliance of the AImaaS ABB of TRUSTEE is investigated with regards to the user-centred, technical, legal, and socio-ethical requirements that were reported in D2.1 (TRUSTEE consortium, 2023). The results of this investigation provide further preliminary insight regarding with the aforementioned requirements addressed during the development and by the implementation of the DPIA and are presented in the following sub-sections. It is worth noting that only the Mandatory, Desirable, and Optional requirements have been considered at this point. Furthermore, only the currently compliant or partially-compliant requirements have been considered. As the implementation of the DPIA progresses during the development of the TRUSTEE Platform and the maturity of the ABB increases, the investigation of the compliance will evolve and further updates and/or enhancements will be provided in upcoming versions of this deliverable, namely D3.2 and D3.3.

User-centred requirements were identified through the UX Research that was conducted under T2.3. These requirements were reported in D2.1 (TRUSTEE consortium, 2023), while they were also utilized in order to extract

technical, functional and non-functional requirements, that were also outlined in D2.1. Table 9 below presents the interplay between user-centred requirements and technical requirements, that is the functional and non-functional requirements that address each user-centred requirement. The user-centred requirement ID is used to reference the respective requirement reported in D2.1 (TRUSTEE consortium, 2023) while the requirement's description is also provided alongside whether the DPIA is currently fully or partially compliant with the user-centred requirement, i.e., indicating whether technical requirements have been identified in order to address the respective user-centred requirement. The ID of the technical requirement, as reported in D2.1 (TRUSTEE consortium, 2023), that is addressing each user-centred requirement is also provided in the last column of the table. Additionally, it is possible than more than one technical requirement addresses a single user-centred requirement.

Table 9: AlmaaS – User-centred requirements

Req id (WP3)	Req. Description	Compliance (Y/N/Partially)	How is addressed
USR-031	TRUSTEE should explain how the AI model process generates the additional knowledge extracted from data	Partially	Req-AIMaaS-FUNC-1 Req-AIMaaS-FUNC-2
USR-052	TRUSTEE should extract knowledge without complex analysis methods and tools	Yes	Req-AIMaaS-FUNC-2 Req-AIMaaS-FUNC-5
USR-053	TRUSTEE should reuse already trained ML and AI models/algorithms	Yes	Req-AIMaaS-FUNC-4 Req-AIMaaS-FUNC-5
USR-059	TRUSTEE should be user-friendly	Yes	Req-AIMaaS-NFUNC-3
USR-062	TRUSTEE should have a user-friendly validation and presentation of data analysis results	Yes	Req-AIMaaS-NFUNC-3 Req-AIMaaS-NFUNC-4
USR-085	TRUSTEE could support improved AI-based analysis methods and tools	YES	Req-AIMaaS-FUNC-4 Req-AIMaaS-FUNC-5
USR-089	TRUSTEE could allow data computations with low computational power	YES	Req-AIMaaS-FUNC-5

USR-092	TRUSTEE could enable validation of AI support decisions	YES	Req-AIMaaS-FUNC-5
USR-114	TRUSTEE could utilize the same algorithms to manipulate different types of data	Partially	Req-AIMaaS-NFUNC-5 Req-AIMaaS-NFUNC-6

Table 10 presents the preliminary results of the investigation regarding the compliance of the HEDF with the technical, functional and non-functional requirements specified in D2.1. The ID of the respective technical requirement is used as a reference to D2.1, while the requirement's description is also provided alongside whether DA complies with the requirement and how. Further updates and/or enhancements regarding the approach presented below for addressing each technical requirement will be provided in upcoming versions of this deliverable, namely D3.2 and D3.3.

Table 10: AIMaaS – Functional and non-functional technical requirements

Req id	Req. Description	Compliance (Yes/No/Partially)	How is addressed
Req-AIMaaS-FUNC-1	The AIMaaS tool should include explainable AI functions	Yes	The xAI subcomponent of the AIMaaS ABB/tool provides explanations for the decisions of the involved AI models. Currently, the XAI subcomponent produces feature importances and presents several visually informative plots.
Req-AIMaaS-FUNC-2	The AIMaaS tool could provide a model for the generation of synthetic data, based on real-world data	Partially	FORTH will provide a number of models that will allow generation of synthetic data for healthcare
Req-AIMaaS-NFUNC-3	The AIMaaS tool should have an easy-to-use and efficient UI	YES	We have designed a User Friendly interface – like the DA UI in order to accommodate the end user friendliness (testes and validation are to be composed)

Req-AIMaaS-FUNC-4	The TRUSTEE platform should store trained AI/ML models for the users to reuse.	YES	The Global Model which occurs as the result of the employed FL scheme, is stored to the TRUSTEE Data Lake,
Req-AIMaaS-FUNC-5	The AIMaaS tool could enable the Model Provider to share self-developed improved AI/ML models for the Consumer and Developer to use.	YES	Via the FL module, TRUSTEE users will be able to share self-developed and locally trained models for improving their performance, based on the privacy-preserving FL paradigm.
Req-AIMaaS-NFUNC-1	The AIMaaS tool should utilize a Docker container to provide the deployment and run of AI/ML models in an easy and fast way.	Yes	Three containers at least will be realized each one for each submodule of this ABB
Req-AIMaaS-NFUNC-2	The TRUSTEE platform should store trained AI/ML models for the users to reuse	Yes	TRUSTEE platform stores trained AI/ML models
Req-AIMaaS-NFUNC-3	The AIMaaS tool should have an easy-to-use and efficient UI	Yes	The UI is following the DA design and is showcased in this chapter
Req-AIMaaS-NFUNC-4	The AIMaaS tool should include functionalities related to user-friendly results presentation	Yes	The UI is following the DA design and is showcased in this chapter showing userfriendliness
Req-AIMaaS-NFUNC-5	The AIMaaS tool could enable the developer to share self-developed improved AI/ML models	Partially	The design of TRUSTEE AI models is based on this requirement but is to be further investigated and current implementation to be improved
Req-AIMaaS-NFUNC-6	The AIMaaS tool could redirect the users of AI models to the Trustworthiness AI Support Design Framework, ensuring the use of trustworthy AI models.	Partially	The design of TRUSTEE AI models is based on this requirement but is to be further investigated and

current implementation to be improved

A preliminary analysis of the compliance with the legal and socio-ethical requirements defined in the D2.1 (TRUSTEE consortium, 2023) is shown in ANNEX 2 and ANNEX 3.

4.4. ARCHITECTURE

4.4.1. HIGH LEVEL

Besides simple AI/ML models that will be included, the AIMaaS will also offer FL models for small fast analytics need in the source. The explainable AI functions will be incorporated into the AIMaaS block, so no external interface was defined. The ABB application view diagram, the sequence diagrams that correspond to the flow of a Consumer running a model and the flow of a Model Provider training and storing their model, as well as the external interfaces definition, follow.

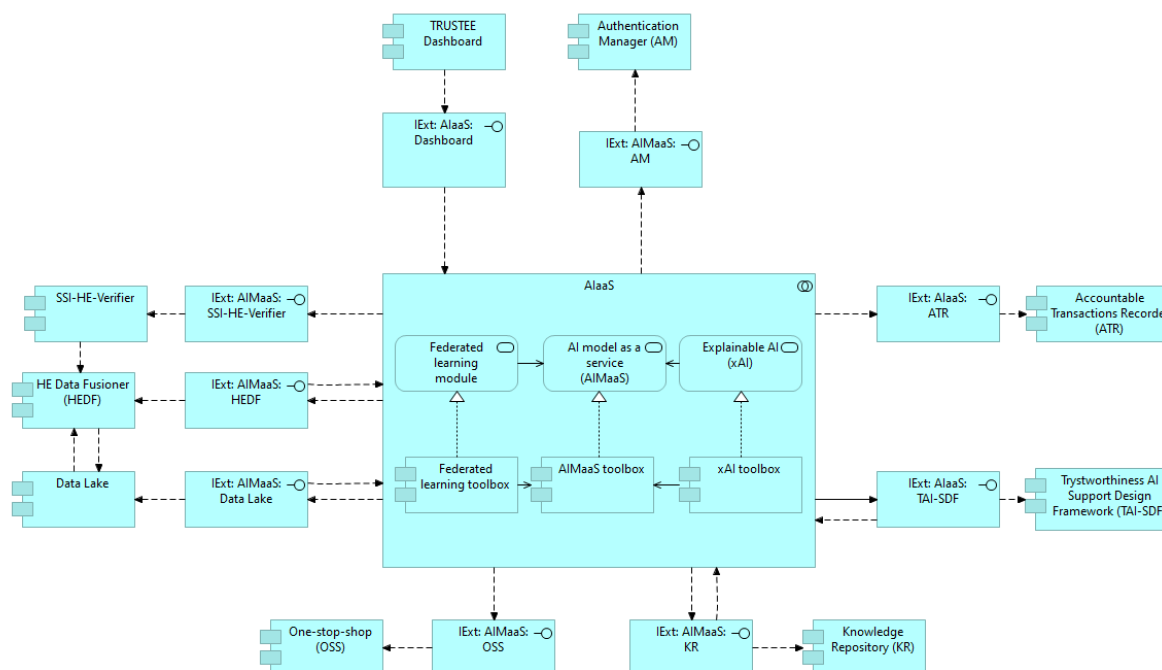
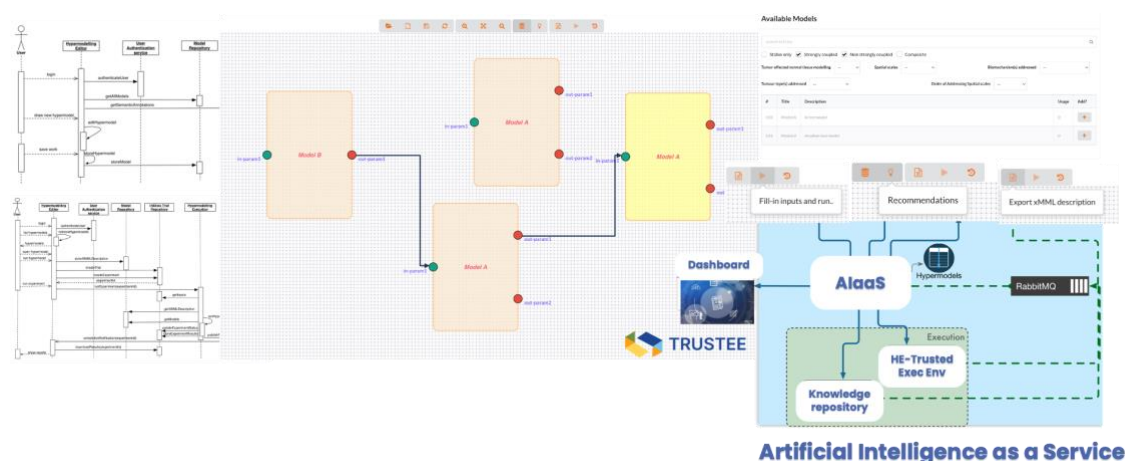
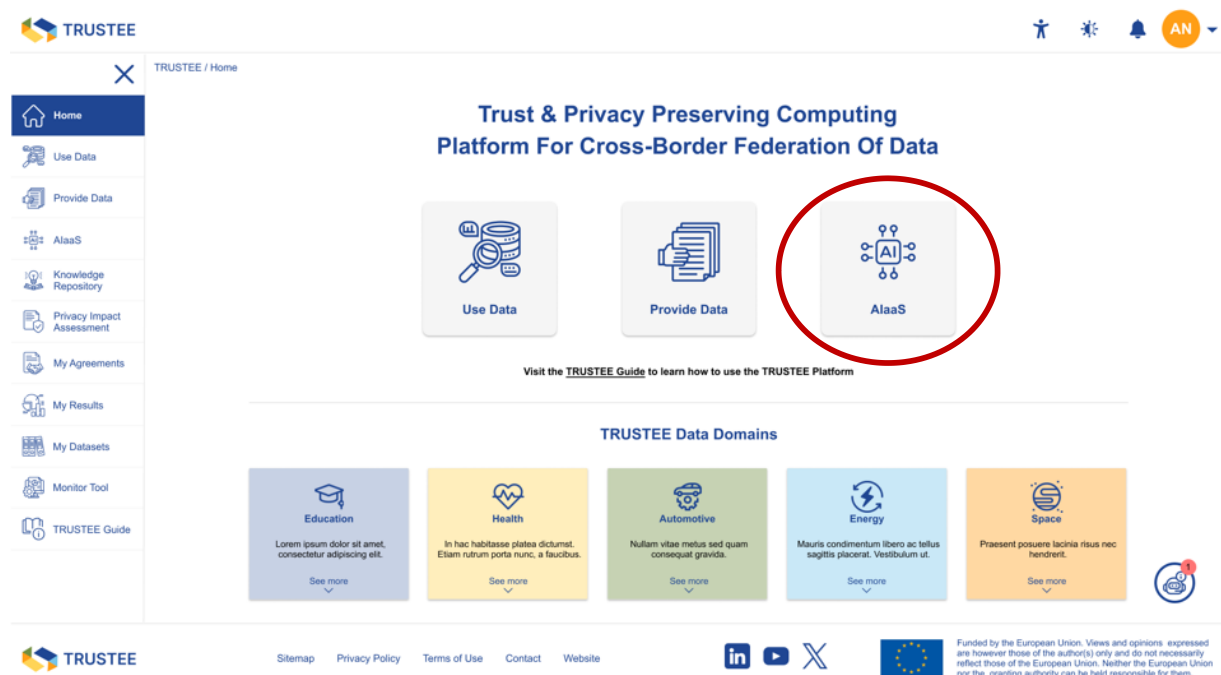


Figure 24: AIMaaS component diagram

4.4.2. LOW LEVEL

AI Model As A Service

The AI Models as a Service building block will enable a plug-n-play approach for the on-boarding of various AI models for data processing. This block will offer functionalities for both Consumers, as well as Model Providers. Consumers will be able to launch the AIMaaS tool directly through the DA ABB, search for datasets of their choice using the SSI-HE framework, search for available AI models, and run a specific AI model on specific data, as a service. On the other hand, the Model Providers will be able to upload to the TRUSTEE platform their pre-trained models, they will have the opportunity to run explainable functions to their trained models and perform a trustworthiness assessment of the model through the TAI-SDF ABB. Finally, they will be able to store their models within the TRUSTEE environment in a form that will enable Consumers to use them in an easy and fast mode. Below we present a mockup of the high-level concept sequence diagrams, high level architecture and front end.



AIMaaS/ FL module

The AIMaaS/FL module will handle all functionalities pertaining to the HE-enabled FL paradigm offered by the TRUSTEE platform to registered FL prosumers. In more detail, the module will define a set of suitable user-selectable attributes for configuring the collaborative FL training mechanism. Moreover, a set of selectable FL-specific aggregation operations will be investigated and implemented, including weighted average for clustered FL, personalized FL, and extended fusion rules incorporating xAI-by-design (i.e., deep unrolling) functionality into the FL pipeline. Additionally, the module will include all required functionalities for tracking and orchestrating the entailed iterative training process of HE-enabled FL. Finally, the module will include all necessary internal and external interfaces for communication with the UI and xAI modules of AIaaS, as well as with other TRUSTEE ABBs, including HEDF (for performing the aggregation operation in the encrypted domain and decrypting the obtained global model), and DL (for storing user-selectable attributes and for accessing the decrypted global after each FL round). An overview of the module is presented in Figure 25.

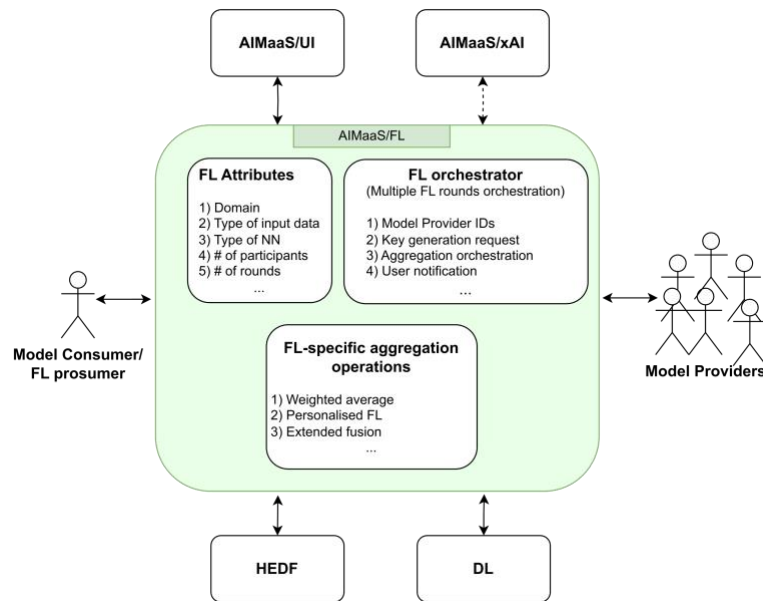


Figure 25. The main interfaces and functionalities of the AIMaaS/FL module

xAI subcomponent

The aim of the xAI subcomponent is to help the end users better understand the ML/DL models that are involved and to explain their predictions. The task of the specific ML/DL model application, e.g. Binary/Multiclass classification, Regression, will be identified and provided to the user. In addition, the main properties and architecture of the ML/DL model, e.g. Decision Tree, Ensemble methods, MLP, CNN, etc. will be analysed and presented. According to the nature of the ML/DL models that are utilized, one or more XAI techniques, e.g. SHAP, LIME will be applied, to explain global/local predictions of the model. The results will be presented by using easily understandable visualizations that will help the end users to understand how the ML/DL model utilizes the input data and how the input features affect the final model prediction.

The current version of the XAI subcomponent which is depicted in the figure below takes a pre-trained model and a dataset as input and produces explanations in the form of feature importances as output. It also generates insightful plots to visualize the results of the explanations.

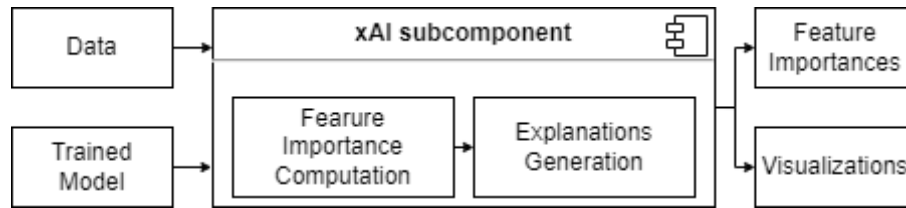


Figure 26: xAI subcomponent Inputs/Outputs

4.5. IMPLEMENTATION

4.5.1. FUNCTIONAL DESCRIPTION

4.5.1.1. AI AS A SERVICE

The purpose of this section is to document the design, the functionality, and the architecture of the AIaaS Inference (AIaaS_i) component. The AIaaS Inference subcomponent of TRUSTEE is the end user software web component for designing, triggering and monitoring the execution, and retrieving the results of the TRUSTEE integrative AIaaS model inference. The tool supports the construction of new complicated workflow of AI-models (already included within TRUSTEE) in a visual and graphical way. It features a “box and arrows” representation of the workflow, where each model is shown as a rectangle and each data exchange link between two models is depicted as a line connecting them. The tool, however, is not an autonomous component in the TRUSTEE platform: it builds upon and integrates with the data lake, knowledge repository, data fusion, SSI-HE, among the others, TRUSTEE component framework. The primary goal is to design new AI workflows from existing data/knowledge/AI models within TRUSTEE in order for the user to browse, search, and retrieve the available models based on their semantics/annotations and metadata-based descriptions. At the level of the **AIaaS** no implementation specific information is required for the models that the users have access to. On the other hand, such information is obviously required in the execution framework in order for the designed workflow to run within a secured and trusted execution environment. Therefore, there’s a tight integration between the different components to collectively implement the design and execution of the TRUSTEE integrative AIaaS.

4.5.1.2. FEDERATED LEARNING

The FL subcomponents handles all required functionalities for performing AI model training via the Federated Learning paradigm, utilizing the TRUSTEE platform. More specifically, this entails the following main functionalities:

- Formation of the group comprising the TRUSTEE users that agree to cooperate in FL-based AI model training.
- Selection of the FL variant to be executed.
- Specification of model architecture and data to be used.
- Configuration of the selected FL training procedure.
- Orchestration of FL training rounds.

4.5.1.3. EXPLAINABLE AI

The main functionality of the xAI subcomponent is to provide coherent, human-understandable explanations of the AI-models that are utilized in TRUSTEE. In more detail, the functionalities that are included in the xAI subcomponent are as follows:

- a. **Model Agnostic Explanations:** The xAI subcomponent employs model agnostic xAI techniques, in order to be able to provide explanations about black-box ML/DL models, which are not specified. This feature adds to the flexibility of the TRUSTEE platform, as the model developers are not limited by a specific only model type that is supported.
- b. **Local Explanations:** This functionality focuses on explaining individual predictions made by the model. To this regard, the xAI subcomponent identifies and quantifies the contribution of each feature in the input data to a specific prediction.
- c. **Global Explanations:** This functionality aims to provide a broader understanding of the model's behavior by ranking the features based on their overall impact on model predictions. This functionality allows users to comprehend which features are generally more critical across the entire dataset, providing valuable insights into the model's decision-making process.
- d. **Explanations of Tabular-data-related Models:** This functionality is tailored for models that handle tabular data, such as structured datasets in a table format. The xAI subcomponent will use adapted techniques to generate insights into how individual features affect the model's predictions for tabular data.
- e. **Visualization Mechanisms:** To facilitate intuitive understanding, the xAI subcomponent incorporates various visualization mechanisms. This includes visualizing feature importance scores, highlighting relevant regions in images, or displaying attributions of each feature's impact on a particular prediction. These visualizations can provide users with a clear and interpretable representation of model behavior, making it easier to understand the reasons behind model predictions.
- f. **XAI-by design methodology:** This functionality provides an efficient and effective framework for building interpretable AI models that are more concise and require less training data to achieve state-of-the-art performance compared to the classical deep learning networks. This stems from the fact that the adopted methodology considers the underlying physical processes and utilize prior domain knowledge of the examined problem.

4.5.2. OPERATIONAL DESCRIPTION

4.5.2.1. AI AS A SERVICE

The users that will be able to interact with the AIMaaS tool including the **consumer** and the **provider**:

- The **model provider** will be able to upload a self-developed and pre-trained AI model to the TRUSTEE platform, to be stored within the TRUSTEE environment and to be used in the future by data consumers.
- The **data prosumer** will be able use TRUSTEE to locally train a FL model through TRUSTEE

The model provider sequence diagram, see figure below showcases the flow of a model provider who wishes to upload a self-developed and pre-trained model on the TRUSTEE platform. As in the data consumer flow, the model provider will enter the TRUSTEE environment through the Dashboard, be redirected to the AIMaaS module and their authentication will be checked by the Authentication Manager. Afterwards, the model provider needs to access the TRUSTEE wiki that resides in the OSS, in order to check the steps that needs to be followed. The first step includes running an xAI function on the trained model. Then, the model provider can start the TAI-SDF module, so they can run a trustworthiness assessment on the model and get a corresponding report back. It is the model provider's responsibility to create all the necessary files in order to run the model in a container. Then, the model

provider needs to send the necessary files (including the docker image, the model description, the trustworthiness assessment report, etc.) to the Data Lake, where they will be stored. The model index of this newly stored model can be sent back to the model provider for future use. Again, as in the case of the data consumer flow, after every action that is performed, the corresponding information is sent to the ATR, so that it can record the related transaction.

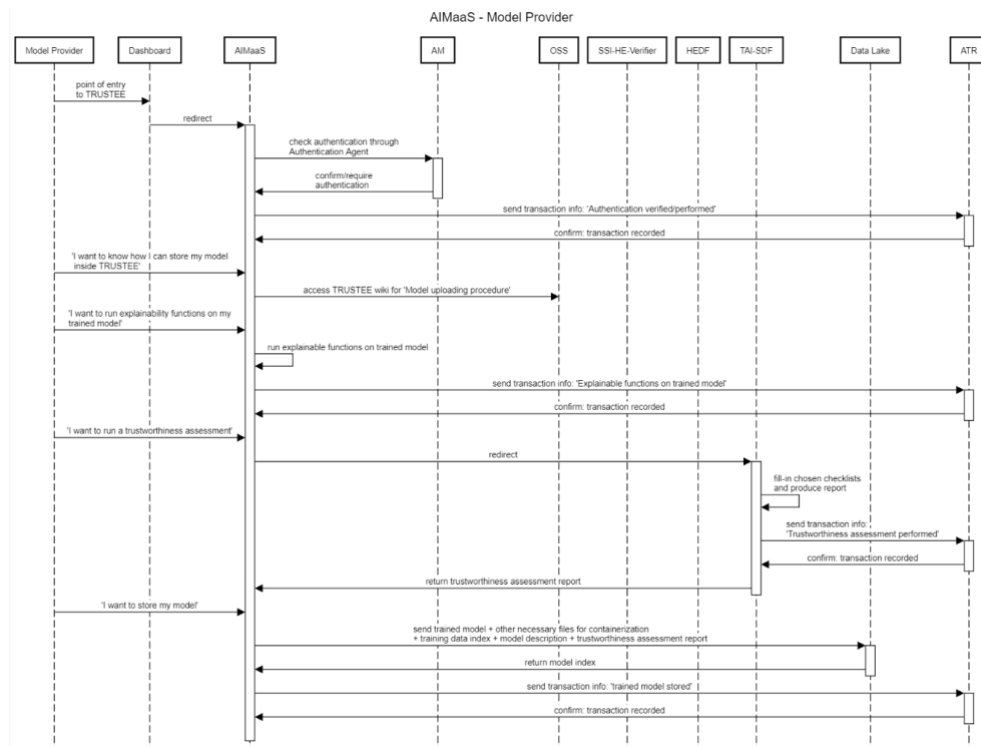


Figure 27: Model Provider Sequence diagram

More than that through TRUSTEE the prosumer case will be that, that describes the case where a user tries to locate a model to perform local FL training and thus needs to be guided through the various ABBs through AIaaS user interface. The sequence diagram below provides details for all steps and also some indicative mockup screenshots of the developed (but not finalized interface are provided).

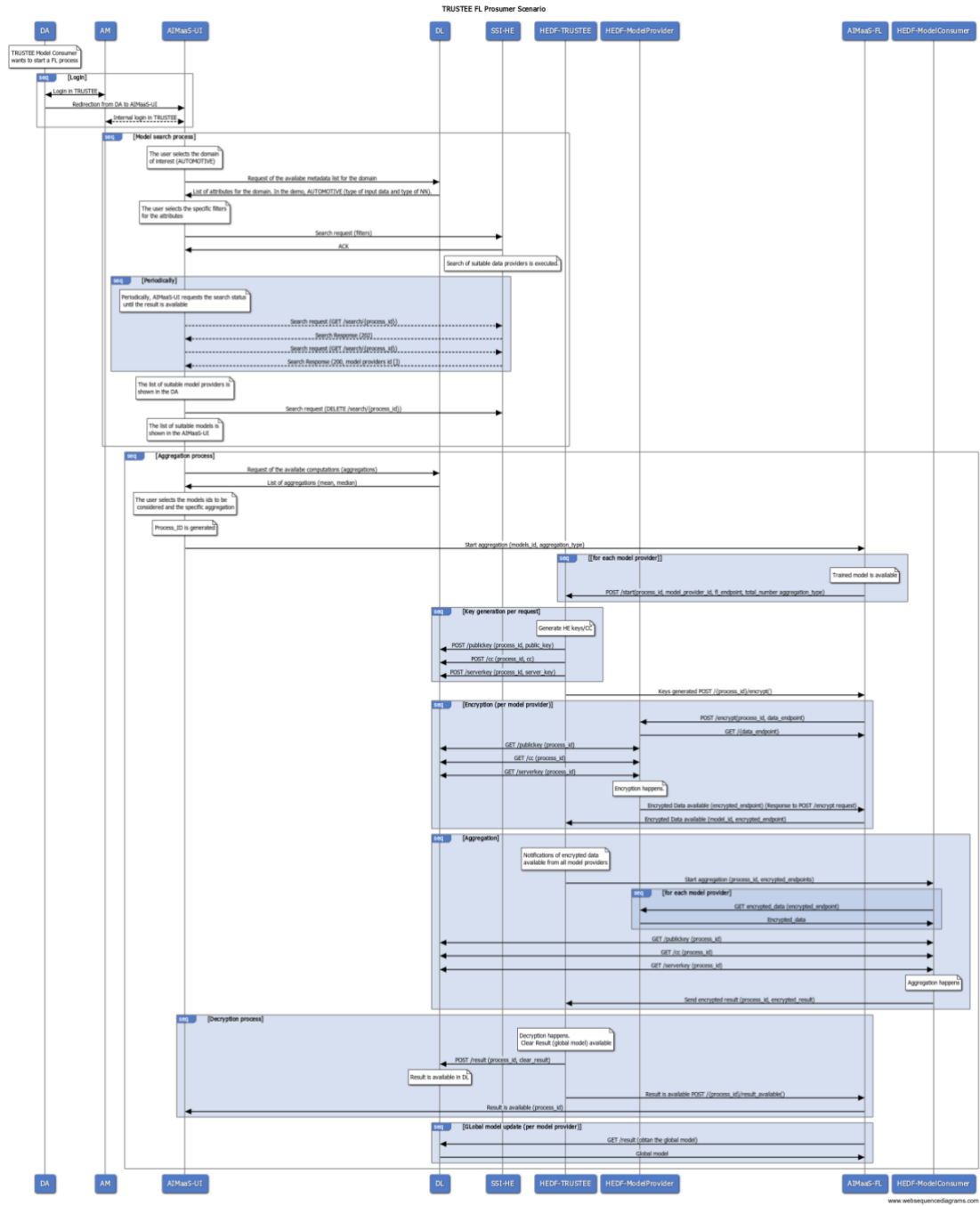


Figure 28 Prosumer use case

- **DL** (IExt:AIMaaS:DL). The AI models will be stored in the DL, so the data consumers will be able to search and choose a model of their choice, while the model providers will be able to store there all the necessary files to deploy and run their model as a service.
- **ATR** (IExt:AIMaaS:ATR). All the transactions performed by any user will be recorded in the ATR.
- **KR** (IExt:AIMaaS:KR). The result of all the inference runs of the AI models will be stored in the knowledge repository (if the Model Provider and potential Data Provider have agreed on this), so that if a future user would want to run the same model on the same data, the result would be retrieved from KR, thus speeding up the process.
- **TAI-SDF** (IExt:AIMaaS:TAI-SDF). The model providers who want to offer their models to the TRUSTEE platform will be able to perform a trustworthiness assessment of their models and receive a respective report.
- **OneSS** (IExt:AIMaaS:OneSS). The model provider will need to access the TRUSTEE wiki, which is part of the OneSS. Thus, they will be guided to follow all the necessary steps, in order to upload their model in the appropriate form, for the consumers to be able to run them. Additionally, several functionalities of the AIMaaS ABB will be exposed via APIs through the OneSS for the Developer Persona to use as part of the incremental strategy of TRUSTEE; However, in this case no data of the Data Providers will be made available and/or AI/ML models trained with the data of the Data Providers. The functionalities to be offered through OneSS will be further investigated in the upcoming months of the project and will be further discussed in next versions of this deliverable, namely D2.2 and D2.3 as well as in upcoming deliverables of WP3 and WP4.

4.5.2.2. FEDERATED LEARNING

The users that will be able to interact with the FL module, will first be required to form an “FL group” of TRUSTEE users that will participate in the HE-enabled AI model training procedure following the FL paradigm.

To this end, the participating nodes, acting as TRUSTEE Model Providers, will enter the TRUSTEE environment through the Dashboard, be redirected to AIMaaS and their authentication will be checked by the Authentication Manager (this functionality is inherited by the AIMaaS ABB). Subsequently, they register their model along with all required metadata, containing information about the model architecture, the input and output data (e.g., labeled/segmented images, labeled/segmented Lidar data, etc.) to the Data Lake, after receiving validation from the SSI-HE ABB.

Upon receiving a “group formation” request, the FL module identifies other model providers with similar model architectures and similar input/output data, based on previous commits, and registers the Group ID (containing all the constituent participants’ info) to the Data Lake. A Model Provider could also join an FL group at a later stage, by issuing a “join group” request to the FL module, along with its model details.

All relevant transactions are registered to the ATR.

These initialization steps are summarized in the sequence diagram shown in Figure 30 below.

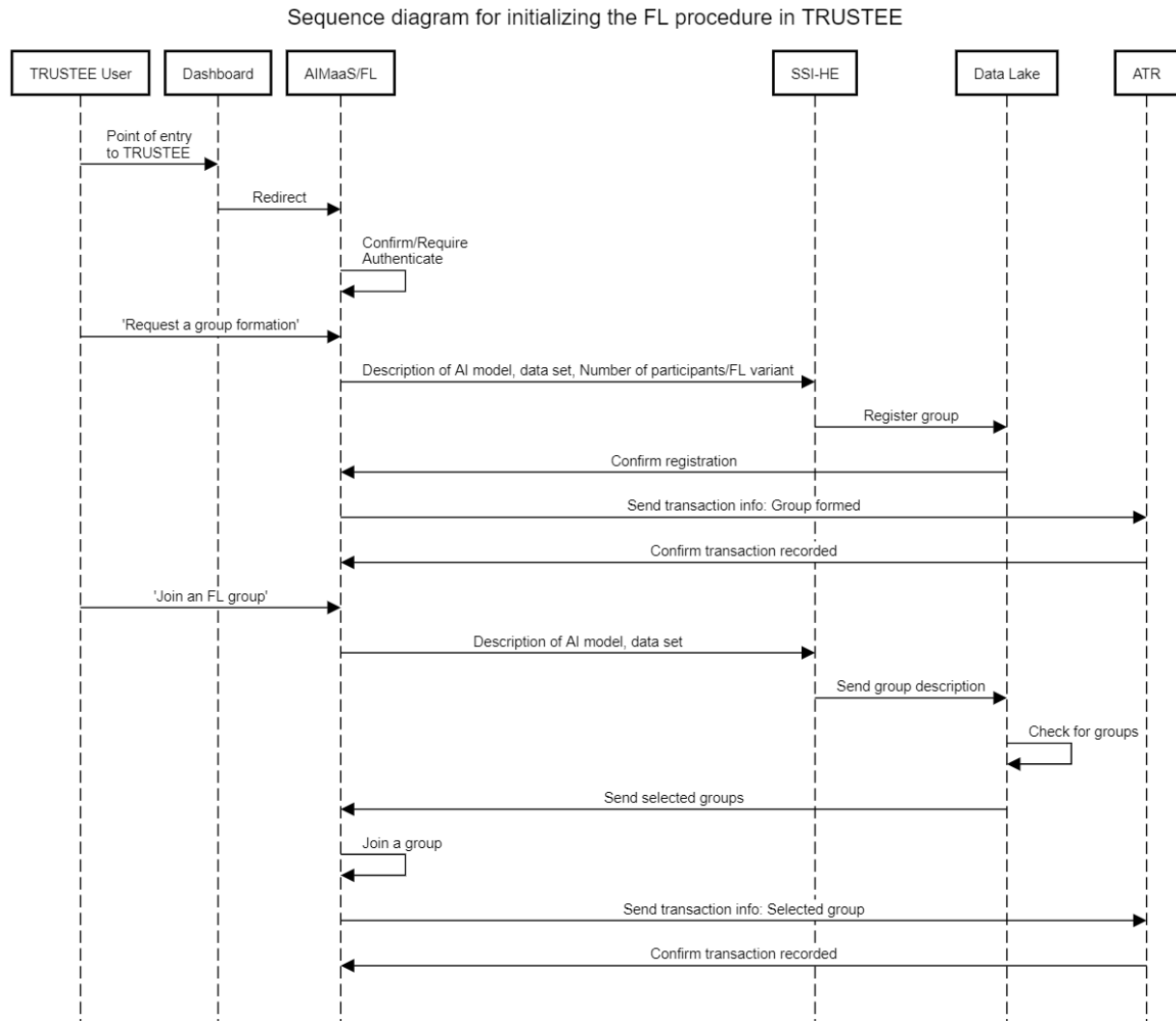


Figure 30. Initialization steps of the FL procedure

Once the potential collaborating group has been identified and registered, the participating users are engaged in collaborative AI model training, assuming interchangeably the roles of **Model Consumer** and **Model Provider**, for a number of predefined Federated Learning rounds. In more detail the executed operations are as follows:

Initially, the group participants train their current version of their local AI model using their respective local (private) datasets producing N updated local models. After receiving a “share local models” request from the FL module, the participants, acting as Model Providers, register their models to the Data Lake, along with the relevant attributes (e.g., number of epochs used in local training, current FL round, etc.) by means of the SSI-HE ABB. After all (updated) local models have been registered to the Data Lake, the FL module send a “local update completed” notification to the group participants. At this point, the HE-enabled aggregation takes place via the HEDF ABB. To this end, each participant, acting as a Consumer, requests a “model aggregation” operation from the FL module, which prompts the latter to initiate a sequence of operations involving the SSI-HE, Data Lake, HEDF, KR, and ATR ABBs. The Consumer receives the (clear/decrypted) updated global model as the final outcome of this sequence. Having this model as a starting point the procedure is repeated for a number of predefined FL rounds.

These steps are summarized in the following Figure 31.

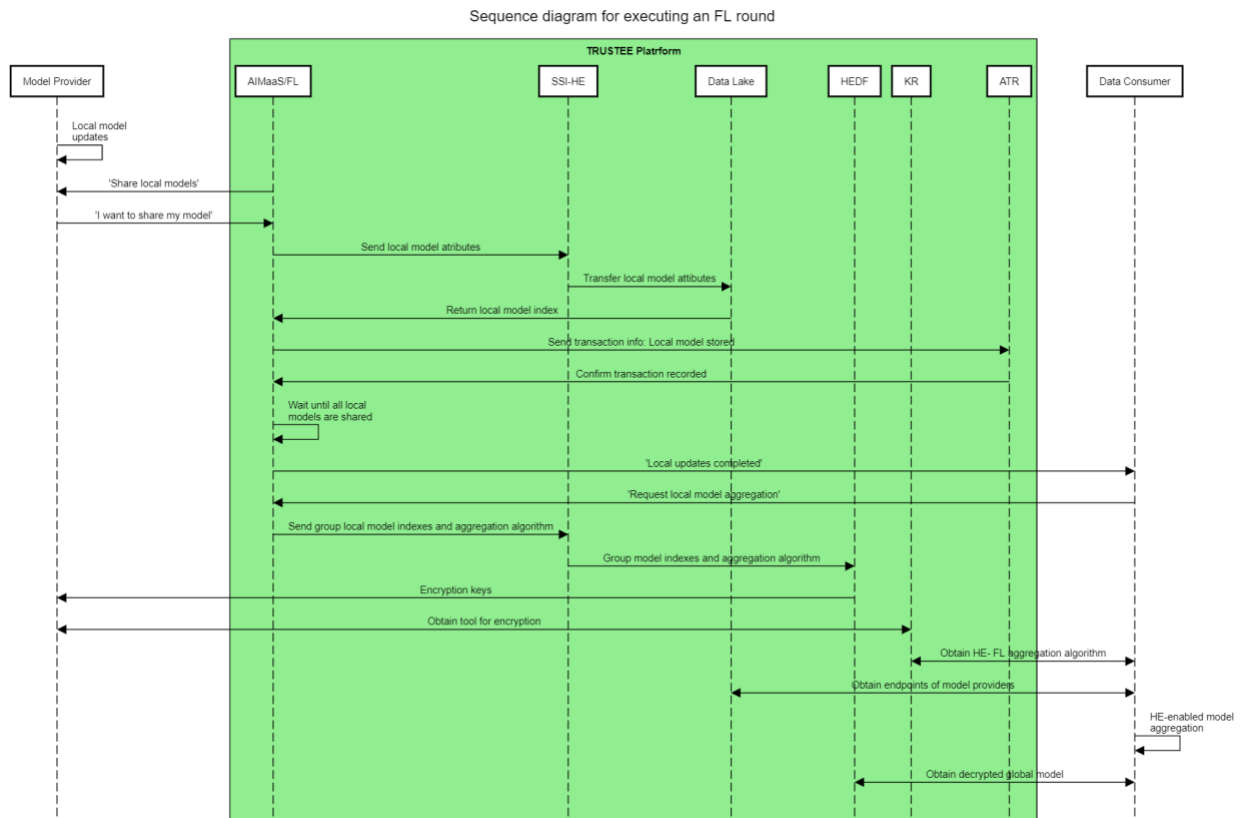


Figure 31. Steps for performing a homomorphically encrypted FL round.

4.5.2.3. EXPLAINABLE AI

The xAI subcomponent functions autonomously within the AIMaaS ABB. It processes input data and a pre-trained model, producing detailed explanations in the form of feature importances. Additionally, the subcomponent generates visual plots to enhance the interpretability of the explanation results. The operational flow of the xAI subcomponent encompasses a series of steps designed to provide users with comprehensive explanations for machine learning model predictions which can be outlined as follows:

- Upon acknowledging a dataset and a pre-trained ML model as input, the xAI subcomponent prepares to generate explanations for the model's predictions.
- Leveraging state-of-the-art explainability techniques, the xAI subcomponent analyzes the input data and model to generate detailed explanations. These explanations manifest as feature importances, shedding light on the factors contributing to specific predictions.
- Recognizing the importance of user-friendly interpretation, the xAI subcomponent goes beyond numerical outputs. It produces visual plots that offer an intuitive representation of the critical features influencing the model's predictions, further enhancing user understanding.

Designed with flexibility in mind, the xAI subcomponent is currently tailored to handle tabular data. Anticipating future needs, the next version will include support for image data, expanding its adaptability to diverse data types.

It is also designed to adapt to various model types. This approach ensures that the subcomponent remains versatile in addressing a wide range of use cases.

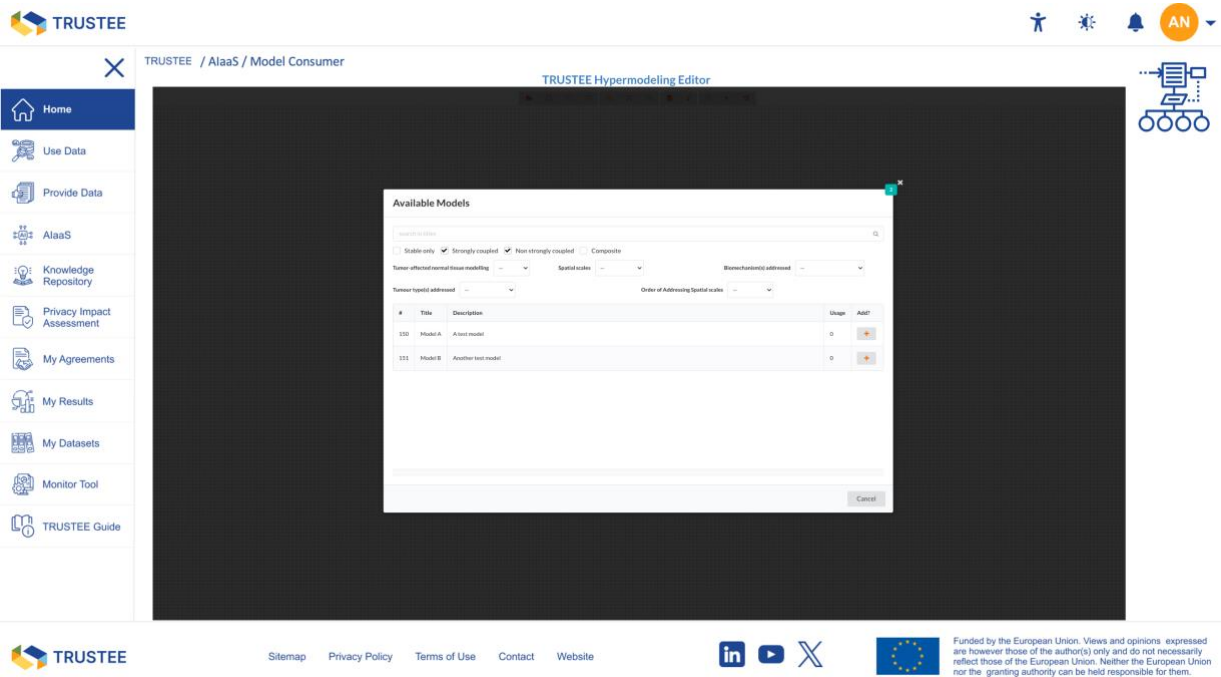
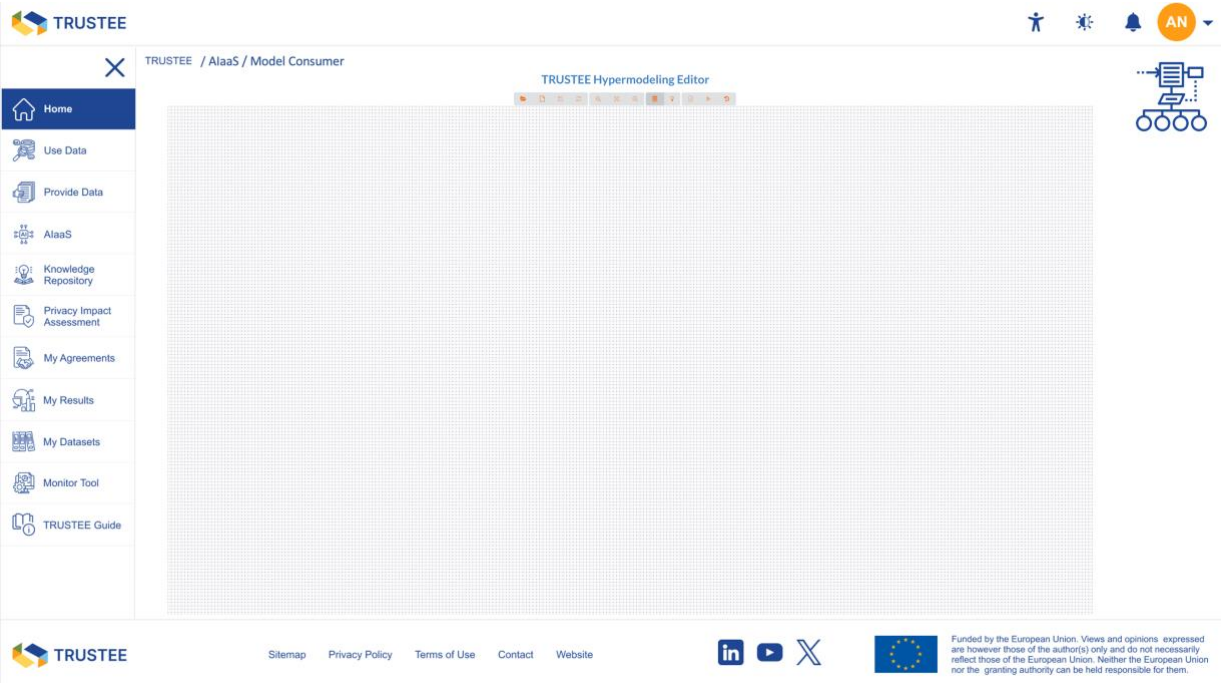
Summarizing, the xAI subcomponent can deliver the generated explanations, both in numerical form (feature importances) and through visual plots, providing users with comprehensive insights into the model's reasoning.

4.5.3. TECHNICAL DESCRIPTION

4.5.3.1. AI AS A SERVICE

The **AIaaS** features a graphical user interface for the visual construction of the models' workflow. The dashboard is responsible for enabling this module and the rationale is that this graphical user interface presents a lower barrier to entry, especially for (ICT) inexperienced researchers. After all, as everyone knows, “a picture is worth a thousand words”, meaning that it is much easier to get a message across using pictures and visual embellishments than it is with words. It is evident that a scripting or compiled programming language can possibly be a lot more expressive and flexible than a visual notation at the cost of a steeper learning curve and the inclusion of more “ceremony” (irrelevant details). There's also a clear trade-off in complexity and flexibility, and the current design choices move a lot of complexity in the platform and the implementation of the various tools, relieving the users of additional cognitive load: the users of the AIaaS need to focus only on the domain specific aspects of the AI-inference building tasks. The provision of such visual interface allows for an additional advantage: there is a shift of focus from the implementation details of each sub-model to the abstract level of general AI workflow of models and the most applicable strategy to use. The concept is that each designed workflow presents a stripped version of a pipeline where any technical artefacts of all implementations are removed, leaving only the domain specific information about its role, data, characteristics, etc. The **AIaaS** therefore with its visual programming interface supports the abstract representation of the models and effectively hides the complexities of the “concrete world”, such as the execution requirements, the management of input and output data, etc. (the “sausage factory” details).

In the graphical representation of the tool, each model is represented by its abstract view and is shown with a minimal set of information that includes only its inputs and outputs, as figure below depicts. The users have of course full access to the whole set of metadata attached to a given model, but in the “drawing canvas” of the tool each model is represented as a rectangular box with a set of input and output ports (circles) that allow the user to connect this model with other models. Each connection between two models represents a flow of data from one output port to one input port of the corresponding models. The underlying paradigm is therefore that of Dataflow, where the workflow is a (directed) graph whose nodes are AI-models and data flows from model to model through the edges of the graph. Each workflow can be either a cyclic or acyclic, depending on the models that participate in it. An acyclic graph does not contain cycles and therefore it can have a serialized execution plan at runtime, i.e. the models can be executed one after the other. The tool supports the combination of models that are strongly coupled, which means that during execution they exchange data through some special “dynamic” ports. These dynamic ports are effectively channels that allow the communication of two concurrently running models and therefore they permit the creation of cycles if, for example, there's a bidirectional dependency between two models where each model sends and receives data through these channels. The tool supports the creation of both cyclic and acyclic hypermodels but not the mixing in the same hypermodel of strongly coupled and “weakly” coupled models. This is due to the fact that the “weakly-” and “stronglycoupled” might have different execution requirements and are handled differently.



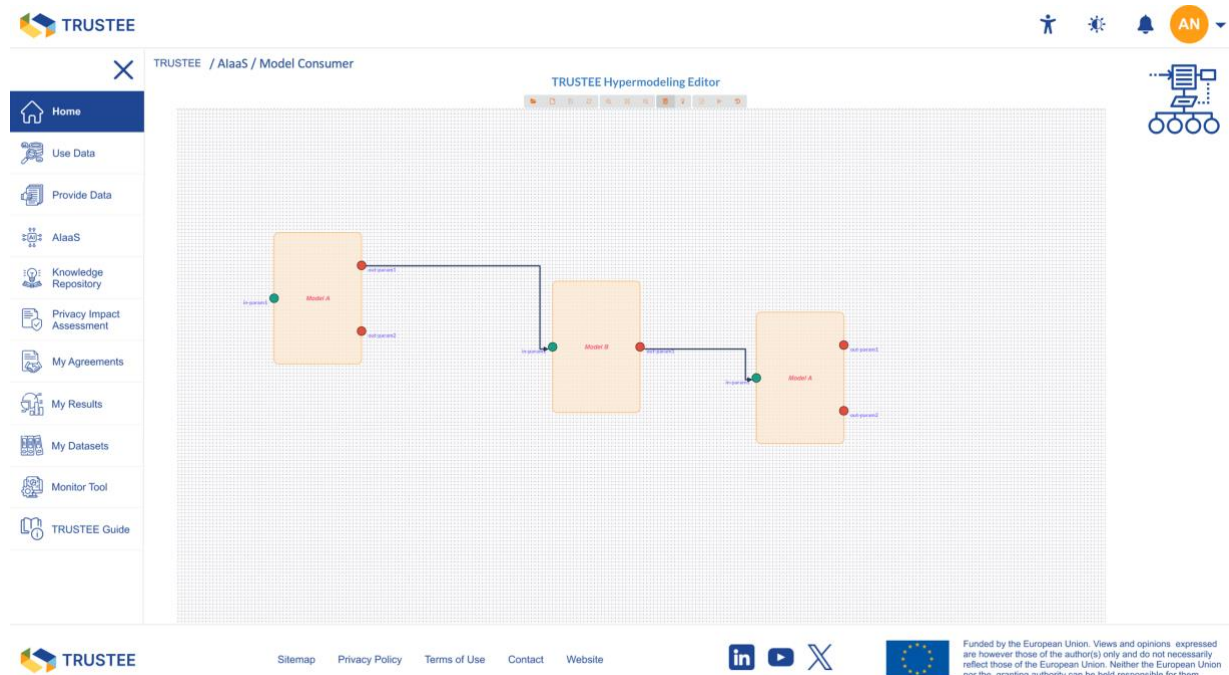


Figure 32: AlaaS TRUSTEE canvas

Management of models

The custodian of models in the platform is the *Data Lake*. This is the component that has the versions of the TRUSTEE AI-models with all the required information that allows their efficient indexing and retrieval. The pipeline workflows designed can be also considered models or hypermodels and therefore the tool is also able to further updates the model repository whenever a new hypermodel is built or an existing one is updated. In addition to the name (title) and the description information, the tool provides the following annotation properties for the new hypermodels: Inputs (multiple); Outputs (multiple); Author (“creator”) information; Publications related information; Domain specific metadata, like temporal and spatial scale information; Versioning information, with links to previous versions (the “parent” version that the current one was based on); Usage and runtime information, e.g. number of times executed, trial context information for each run, etc. Whenever such information is sent to the Data lake it is encoded in the form that the repository supports and uses the agreed coded values (for example, the data types for the inputs and outputs).

Some remarks on the view of the information managed by the tool are in order:

- A workflow/hypermodel is created by a user during her/his interaction with the AlaaS tool.
- Each workflow/hypermodel can have multiple versions, where each version builds upon a previous one (the “parent” version)
- Each workflow/hypermodel bears some descriptive metadata, like the title, description, and creation date, but in fact this information can change from one version to another.
- In addition to the descriptive metadata, each workflow/hypermodel version consists of its specific set of models.

- The models that collectively comprise a given version carry all the annotations managed by the data lake, e.g. the names, data types, and default values of their inputs and outputs.

4.5.3.2. FEDERATED LEARNING

In this section, the implementation details of the current, standalone version of the FL module is presented in order to provide a proof-of-concept demonstration of the envisioned functionality of the module, as well as the benefits that arise when such functionality is provided by the TRUSTEE platform. At the next stages of TRUSTEE, this implementation will be further elaborated by incorporating the desired functionalities by other TRUSTEE ABBs as indicated in Figure 30, Figure 31 in which the sequence diagrams that describe the required steps and interactions within TRUSTEE, are depicted.

In Federated Learning (FL), every client conducts model training within its own premises, utilizing a fully dockerized environment. The technology of containers represents a lightweight and executable software unit that bears application code and its associated dependencies, including binary code, libraries, and configuration files. This packaging enables seamless deployment across diverse computing environments. Frequently dubbed "operating system virtualization," containers can execute software, microservices, or entire applications with efficiency and flexibility.

We used Docker for implementing our containerization strategy. More particularly, we used the Docker Compose Tool for implementing multi-container Docker applications. The tool only needs a YAML configuration file for deploying several containers (see example in Figure 33).

```
version: "3.5"

services:
  foxglove:
    restart: always
    # build: ./foxglove
    image: ghcr.io/jpc/studio:latest
    command: caddy run
    volumes:
      - ../../carlafox/Caddyfile:/src/Caddyfile:ro
      - ../../carlafox/foxglove_layout.json:/src/foxglove_layout.json:ro
    network_mode: host

  ros-bridge:
    restart: always
    init: true
    image: ghcr.io/collabora/ros-bridge:latest
    network_mode: host

  # ports:
  # - "9090"
  #
  # deploy:
  #   resources:
  #     reservations:
  #       devices:
  #         - capabilities: [gpu]

  carla:
    restart: always
    init: true
    image: carlasim/carla:0.9.12
    network_mode: host

    command: ./CarlaUE4.sh -RenderOffScreen -nosound
    # ports:
    # - 127.0.0.1:2000-2002:2000-2002
    #
    # deploy:
    #   resources:
    #     reservations:
    #       devices:
    #         - capabilities: [gpu]

  lego-loam:
    build: ../../LeGO-LOAM-BOR/docker
    restart: always
    image: lego-loam
    command: source ~/catkin_ws/devel/setup.bash && roslaunch lego-loam-bor run.launch lidar_topic:=/carla/ego_vehicle/lidar
    volumes:
      - /home/chragnago/repos/LeGO-LOAM-BOR:/root/catkin_ws/src/LeGO-LOAM-BOR:ro
```

Figure 33 Example of docker-compose.yaml file. It starts several containers including a Carla simulator, a visualization tool and a SLAM algorithm.

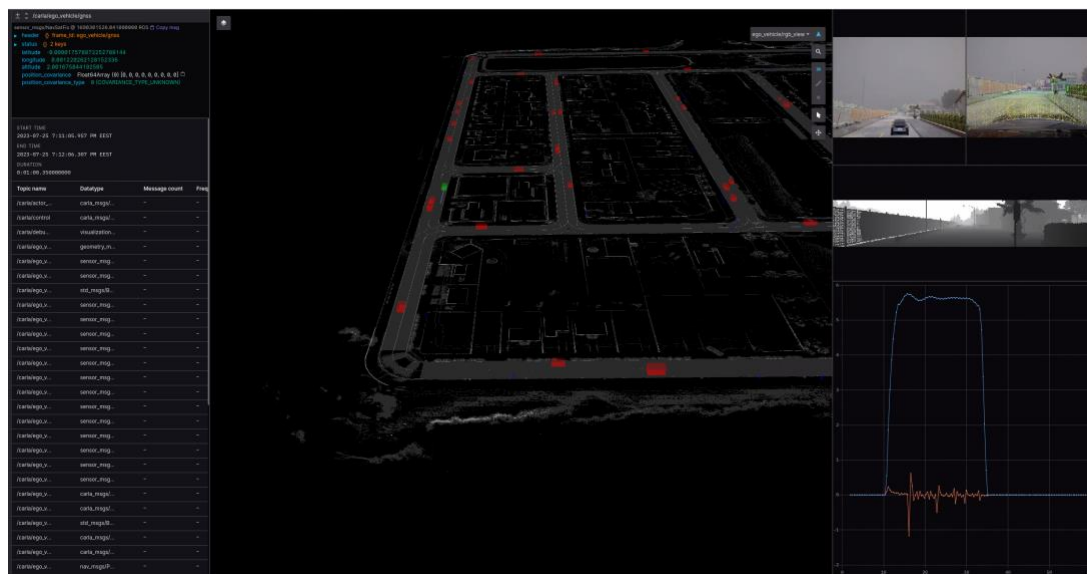


Figure 34 Example of the visualization tool

4.5.3.3. EXPLAINABLE AI

The XAI subcomponent adds an explainability layer after the generated black-box models, which helps users deal with the transparency of machine learning models. By providing comprehensive explanations for complex model decisions, the XAI subcomponent enhances interpretability and fosters trust, enabling users to make informed and reliable decisions based on the model's insights.

The XAI subcomponent focuses on applying interpretability methods to trained models, specifically focusing on post-hoc techniques. By adopting a model-agnostic approach, these methods can be seamlessly applied to any ML model, irrespective of its type (e.g., linear, tree-based, deep neural network), thereby offering remarkable flexibility. In scenarios where the model type is known, model-specific implementations are employed for efficient computation.

This versatile subcomponent's strength lies in its capability to harness both local and global XAI methods, rendering comprehensive explanations. Through local XAI methods, it unravels the decision rationale behind individual predictions, enabling a clear understanding of each model's choices. Concurrently, global XAI methods shed light on the overall behavior of the machine learning model, providing a holistic view of its performance and tendencies. By combining these approaches, the subcomponent significantly enhances interpretability, empowering users to grasp the model's functioning at both granular and aggregate levels.

Currently, the XAI subcomponent produces feature importances as a result of its interpretation methods, augmenting the users' ability to make informed and reliable decisions based on the model's insights.

The main methods that are supported by the xAI subcomponent and have been implemented so far are the following:

Local Interpretable Model-agnostic Explanations (LIME)

LIME is an approach for explaining the predictions of any model in a faithful and interpretable way (Ribero 2016). LIME is model-agnostic and thus treats the original model as a black box. It works by learning an interpretable model locally around the prediction. Overall, LIME provides an explanation of a black box model's prediction for a selected instance by giving textual or visual indications.

Explanations must be interpretable and locally faithful. The first one means that they must provide qualitative understanding between the input variables and the response and the second one that they must correspond to the behaviour of the model in the vicinity of the instance that is being predicted.

The process followed by LIME can be described as follows: After selecting an instance, samples are drawn uniformly at random and then are weighted based on their proximity to the instance of interest. That means that the sampled instances can be both in the vicinity of the instance with a high weight and far from the instance with a low weight. For every sample, the original black-box model's prediction is obtained and used as a label. Lastly, an interpretable model is trained on the new dataset with the samples and their associated labels. The explanation of the prediction of the original black-box model is obtained by interpreting the local model.

The explanation is given by:

$$\xi(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

In order to ensure both interpretability and local fidelity, it is required to minimize the fidelity function $L(f, g, \pi_x)$ which measures how unfaithful model g is in approximating the original model f in the locality defined by the proximity measure π_x while keeping the model complexity $\Omega(g)$ low. Model g can be any interpretable model, like linear model or decision tree.

DeepLIFT (Deep Learning Important Features)

DeepLIFT is a method designed to explain the prediction of deep learning models by calculating the contributions of the network neurons to each input feature (Avanti 2017). This is achieved by comparing the activation of each neuron with its reference activation and assigning contribution scores based on the difference. This method takes into consideration both positive and negative contributions, which makes it possible to uncover dependencies that can be neglected by other methods.

By calculating the contributions using a difference from reference method, DeepLIFT can propagate importances even when the gradient is close to zero value, avoiding the saturation problem that can occur in sensitivity analysis applied on DNNs. The reference activation for each neuron can be chosen in various ways, with a common approach being the activation for a neutral input. The contribution of each feature is then calculated as the sum of the differences across all relevant neuron activations. The ability to calculate the contribution scores in a single backward pass makes this method computationally effective.

SHAP (SHapley Additive exPlanations)

SHAP is a method for explaining predictions of machine learning models (Lundberg 2017). The SHAP method is based on the idea of Shapley values from game theory, which describe the importance of each player in a cooperative game. In machine learning, the "players" are the features and the "game" is the prediction of the model. Using SHAP, we can quantify the contribution of each feature to the model predictions.

To determine the influence of each feature on a specific prediction, SHAP values can be calculated by examining the effect of each feature on the prediction, while considering the relationships between the features. This is done

by first finding the average output for every possible combination of feature values and comparing it to the model's output with the inclusion of the specific feature in question. The resulting difference represents the feature's contribution, or its "fair share," to the prediction given the presence or absence of other features. The SHAP values for all features can be added together to obtain the model's output for the specific prediction.

Due to the large computational effort that is needed, the exact computation of SHAP values is challenging. For this reason, methods that combine SHAP with other XAI techniques have been proposed in order to efficiently approximate the SHAP values. The following SHAP based methods are supported:

- KernelSHAP, is a model-agnostic method that combines LIME and SHAP and calculates the SHAP values by using a weighted linear regression.
- TreeSHAP, as a fast alternative method for the calculation of SHAP values of tree-based machine learning models.
- DeepSHAP is a method that combines SHAP and DeepLIFT, designed specifically to explain predictions of deep learning models. The SHAP values are calculated by using the gradients of the model's outputs with respect to the input features.

The XAI subcomponent serves as a crucial tool in enhancing model interpretability. It accepts pre-trained models and data as inputs and efficiently generates explanations in the form of feature importances, alongside relevant and insightful visualizations. The XAI process follows a series of streamlined steps to ensure effective interpretability. These steps include: a) data and model loading, b) feature importance computation, c) explanation generation and d) visualization.

It starts by loading the input data and a pre-trained model into the XAI subcomponent. Subsequently, in the scenario where the SHAP method is utilized, a SHAP explainer is thoughtfully tailored based on the model type. LinearSHAP, TreeSHAP and DeepSHAP methods are utilized for linear models, tree-based models, and deep neural networks respectively. KernelSHAP is the method of choice for model-agnostic scenarios. The subsequent step involves the efficient computation of SHAP values. These values are instrumental in providing valuable insights into feature contributions, forming the bedrock for generating comprehensive and interpretable explanations. Leveraging the computed SHAP values, the XAI subcomponent generates clear and comprehensive explanations, catering to both individual predictions and the overall model behavior. By associating SHAP values with their corresponding features, the subcomponent accurately identifies feature contributions, thereby enhancing the interpretability of the model's decision-making process. In the final step, the XAI subcomponent presents several visually informative plots to showcase the generated explanations. These plots encompass diverse insights, including feature importances, feature importances with feature effects, feature attributions visualized as "forces", SHAP explanations clustered by explanation similarity, feature dependence, and feature dependence with interaction visualization. Through this rich visualization, users gain a holistic understanding of the model's behavior, the significance of individual features, and the intricate interplay among them, fostering transparency and trust in the model's decision-making process. The available visualizations that are currently supported are the following:

SHAP Force plot

The "force plot" is an explanation visualization tailored for a specific prediction in a machine learning model. It represents feature attributions (SHAP values) as "forces" that influence the model's prediction compared to a baseline, which is typically the average of all predictions. Each feature value acts as a force, either increasing or decreasing the prediction from the baseline. SHAP values appear as arrows that push the prediction up (when positive) or down (when negative). The forces balance each other at the actual prediction of the data instance, offering an intuitive understanding of how each feature contributes to the model's decision for that particular prediction.

SHAP explanations clustered by explanation similarity

SHAP explanations, clustered by explanation similarity, serve as a powerful visualization to understand explanations for an entire dataset. The plot involves clustering the SHAP values of each instance to identify groups of similar instances. It consists of horizontally stacked force plot explanations, rotated by 90 degrees, which are aligned based on their clustering similarity. Each position on the x-axis corresponds to an instance in the data, and the SHAP values are represented using red for increasing the prediction and blue for decreasing it. This visualization offers a comprehensive view of how similar instances contribute to the model's predictions, facilitating a deeper understanding of the model's behaviour across the dataset.

SHAP Feature Importance Plot

The SHAP Feature Importance Plot is a visualization that showcases the ranking of features based on their importance in a ML model's predictions. It is designed to provide insights into the relative influence of each feature on the model's outcomes. In the plot, the features are sorted in descending order based on their importance. The y-axis represents the features, and the x-axis denotes the corresponding SHAP values. Each feature is associated with a bar, and the length of the bar reflects the feature's importance, with longer bars indicating higher significance. This plot serves as a valuable tool for identifying which features have the most impact on the model's decision-making process. It allows users to prioritize and focus on the most influential features, facilitating better understanding and interpretation of the model's behaviour.

SHAP Summary Plot (Feature Importance with Feature Effects)

The SHAP summary plot is a powerful visualization that combines feature importance with feature effects. Each point in the plot represents a SHAP value for a feature and an instance. The position on the y-axis corresponds to the feature, and on the x-axis, it represents the SHAP value. The color indicates the feature's value from low to high. Overlapping points are jittered along the y-axis to observe the distribution of Shapley values per feature. The features are ordered based on their importance. This plot provides a detailed view of the feature effects on predictions, giving deeper interpretability to the model.

SHAP Dependence Plot (Feature Dependence)

While the summary plot offers initial insights into the relationship between feature values and their impact on predictions, to understand the exact relationship form, one needs to refer to SHAP dependence plots. The SHAP dependence plot is a global interpretation plot that simplifies model understanding for a specific feature. It employs a scatter plot, where the x-axis represents the feature value, and the y-axis corresponds to the SHAP value. To enrich the plot further, feature interactions can be highlighted, providing additional insights into the complex relationships between features and their impact on predictions.

SHAP Interaction Plot (Feature Dependence with Interaction)

The SHAP Interaction plot is an insightful visualization that offers a feature dependence plot with a focus on interaction effects. This plot showcases the combined effect of features after considering their individual effects. By examining the interaction effect, we gain a deeper understanding of how the interplay between features impacts the model's predictions beyond their standalone contributions. The SHAP Interaction plot thus provides valuable insights into the complex relationships among features, further enriching the interpretability of the machine learning model.

SHAP Waterfall Plot

The "waterfall plot" provides a visual explanation for a specific prediction by illustrating the contribution of each feature in shifting the prediction from the base value to the actual model output. The base value represents the average model output over the training dataset. In the plot, the y-axis represents the features, and the x-axis depicts

the contribution or impact of each feature on the prediction. Features that push the prediction higher are displayed in red, while those that push it lower are shown in blue. The plot visually demonstrates how the various features collectively influence the model's decision for the specific prediction.

Regarding the input and output of the XAI subcomponent, it requires as input a pre-trained model, and a test dataset containing the instances to be explained. The output of this subcomponent comprises two important elements: a) explanations in the form of feature importances and b) various visualizations that enable users to gain a deeper understanding of the model's behaviour.

XAI-by-design functionality

The XAI-by design functionality provides an efficient and effective methodology that builds upon the model-based deep learning theory, known as deep unrolling methods (Monga 2021). Along with their high interpretability, the derived deep unrolling models are more concise, requiring less training data and number of parameters to achieve state-of-the-art performance compared to the classical deep learning networks.

According to the literature, this methodology is especially suited for addressing inverse problems such as denoising, deblurring and super-resolution, which can assume the following form.

$$\mathbf{X}^* = \underset{\mathbf{X}}{\operatorname{argmin}} \left(\frac{1}{2} \|\mathbf{Y} - \mathbf{S}\mathbf{X}\|_F^2 + \mu R(\mathbf{X}) \right)$$

where the desired signal $\mathbf{X} \in \mathbb{R}^{C \times M}$ is acquired via the transformation $\mathbf{S} \in \mathbb{R}^{C \times C}$ as a degraded signal $\mathbf{Y} \in \mathbb{R}^{C \times M}$, and $c < C$. The cost function consists of a data fidelity term, and an unknown learnable regularizer $R(\cdot)$ that enforces some prior on the solution, trained to capture the structure of the desired signal $\mathbf{X}$. Also, μ denotes a regularization parameter.

The methodology consists of two main steps, namely, (a) the selection of an effective iterative algorithm for solving the minimization problem, and (b) the application of the deep unrolling approach which unrolls a small number of iterations of the selected iterative algorithm to derive a deep network. In more detail, In the first step of the methodology, an appropriate iterative solver is employed. Some notable choices of such algorithms are (a) Gradient descend (for differentiable cost functions), (b) Proximal gradient descend (for non-differentiable cost function), (c) Alternating Direction Method of Multipliers (ADMM), and (d) Half-Quadratic Splitting (HQS). Finally, in the second step, K iterations of the selected algorithm are unrolled using the deep unrolling approach. Each iteration (out of the K) corresponds to a layer of the (eventually) derived deep network.

4.5.4. TECHNICAL DETAILS

4.5.4.1. AI AS A SERVICE

The **AIaaS** therefore with its visual programming interface supports the abstract representation of the models and effectively hides the complexities of the “concrete world”, such as the execution requirements, the management of input and output data, etc. (the “sausage factory” details). In the graphical representation of the tool, each model is represented by its abstract view and is shown with a minimal set of information that includes only its inputs and outputs, as figure below depicts. The users have of course full access to the whole set of metadata attached to a given model, but in the “drawing canvas” of the tool each model is represented as a rectangular box with a set of input and output ports (circles) that allow the user to connect this model with other models. Each connection between two models represents a flow of data from one output port to one input port of the corresponding models. The underlying paradigm is therefore that of Dataflow, where the workflow is a (directed) graph whose nodes are AI-models and data flows from model to model through the edges of the graph. Each workflow can be either a cyclic

or acyclic, depending on the models that participate in it. An acyclic graph does not contain cycles and therefore it can have a serialized execution plan at runtime, i.e. the models can be executed one after the other. The tool supports the combination of models that are strongly coupled, which means that during execution they exchange data through some special “dynamic” ports. These dynamic ports are effectively channels that allow the communication of two concurrently running models and therefore they permit the creation of cycles if, for example, there’s a bidirectional dependency between two models where each model sends and receives data through these channels. The tool supports the creation of both cyclic and acyclic hypermodels but not the mixing in the same hypermodel of strongly coupled and “weakly” coupled models. This is due to the fact that at the “weakly-” and “stronglycoupled” might have different execution requirements and are handled differently.

Management of models

The custodian of models in the platform is the *Data Lake*. This is the component that has the versions of the TRUSTEE AI-models with all the required information that allows their efficient indexing and retrieval. The pipeline workflows designed can be also considered models or hypermodels and therefore the tool is also able to further updates the model repository whenever a new hypermodel is built or an existing one is updated. In addition to the name (title) and the description information, the tool provides the following annotation properties for the new hypermodels: Inputs (multiple); Outputs (multiple); Author (“creator”) information; Publications related information; Domain specific metadata, like temporal and spatial scale information; Versioning information, with links to previous versions (the “parent” version that the current one was based on); Usage and runtime information, e.g. number of times executed, trial context information for each run, etc. Whenever such information is sent to the Data lake it is encoded in the form that the repository supports and uses the agreed coded values (for example, the data types for the inputs and outputs).

Some remarks on the view of the information managed by the tool are in order:

- A workflow/hypermodel is created by a user during her/his interaction with the AIaaS tool.
- Each workflow/hypermodel can have multiple versions, where each version builds upon a previous one (the “parent” version)
- Each workflow/hypermodel bears some descriptive metadata, like the title, description, and creation date, but in fact this information can change from one version to another.
- In addition to the descriptive metadata, each workflow/hypermodel version consists of its specific set of models.
- The models that collectively comprise a given version carry all the annotations managed by the data lake , e.g. the names, data types, and default values of their inputs and outputs.

4.5.4.2. FEDERATED LEARNING

Simulation Framework

For generating synthetic data, we have adopted a simulation framework that combines three sub-systems, namely, a network simulator, a traffic simulator, and a game engine-based simulator. More specifically, we have selected CARLA as the component responsible for simulating physics phenomena and rendering. Additionally, Artery V2X Simulation framework, which is built on top of the OMNET++ framework, was our choice for simulating network

communications and, in particular, V2X communications. Finally, we have chosen SUMO for the control and the coordination of the simulated entities.

Datasets

The dataset that is used for training, could either be a private dataset owned by the client, which may consist of synthetic or real data, or an openly available dataset like KITTI, Waymo, and others. In the context of our experiments, we produced separate datasets for each client participating in FL using the Carla simulator.

ML Models

The models that are trained using the FL module as well as their performance evaluation, are the following:

- *SuperPoint*: A DL-based feature extractor and was used in the frontend of a SLAM algorithm,
- *Super Resolution DNN*: A model that improves the resolution of a low-resolution 3D LiDAR sensor,
- *Yolo7*: A 2D object detector,
- *PointPillars*: A 3D object detector,

KalmaNet: A real-time state estimator that learns from data to carry out efficient Kalman filtering for the localization of autonomous vehicles.

4.5.4.3. EXPLAINABLE AI

Table 11: xAI subcomponent technical details

DEVELOPMENT	
Type (Software/Hardware)	Software
Operating Systems	Windows, Linux (Ubuntu)
Database and databases tools	-
GUI	No
Programming language	Python
Development Environment	Visual Studio Code, Jupyter notebooks
INSTALLATION AND DEPLOYMENT	
Software Requirements	Docker, Python
Hardware Requirements	At least 16GB RAM
Containerization	Docker
Communications	

4.6. VALIDATION

4.6.1.1. AI AS A SERVICE

The service is validated through different scenarios in TRUSTEE.

4.6.1.2. FEDERATED LEARNING

PoC in the automotive domain: applying FL for improving DL-based feature extraction on SLAMs

In this section, the validation of the FL module functionality is performed by demonstrating the gains that are achieved by incorporating an FL mechanism in the TRUSTEE platform and providing it to end-users. For the validation, testing scenarios from the automotive domain are devised.

In Figure 35, the output of a deep unrolling (DU) model for LiDAR super-resolution is employed as the input in a LiDAR SLAM approach. Note that such a DU model can be provided via the XAI-by design module of the AIMaaS ABB. Here, the DU model is learned among several participating vehicles that collaborate via FL. In the Figure, the LiDAR SLAM approach employs the learnt DU model acquired after 6, 24, and 50 rounds of FL. As observed, the error is improved as the number of FL rounds increases. This result provides evidence on the benefits that end-users may have via the FL functionality provided by the TRUSTEE platform.

Finally, in Figure 36, the impact of FL on improving DL-based feature extraction on SLAMs is presented. In more detail, similarly to the previous experimental scenario, the employed DL model is learnt by several participating vehicles via FL. It is observed that the performance is improved when the number of FL rounds increases before the DL model is employed. Here, the performance is reflected by the increased number of feature matches and the improved localization error.

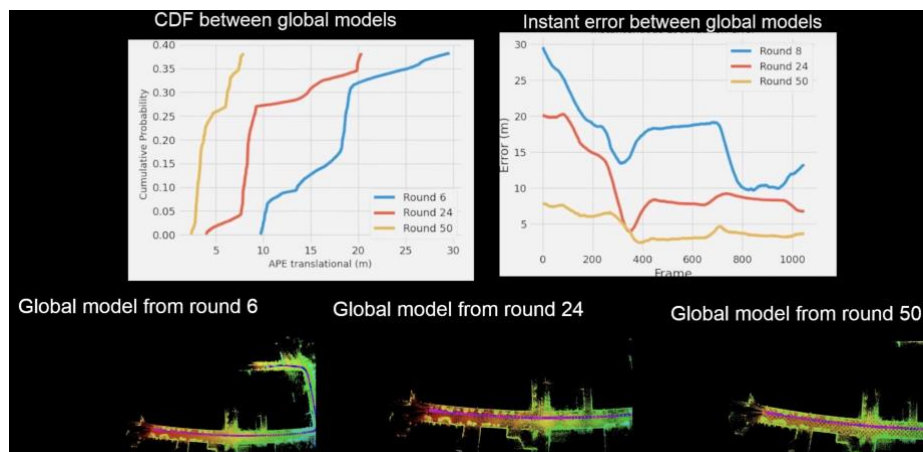


Figure 35 Impact of FL on improving raw data quality using DU models and the benefits on Lidar SLAM

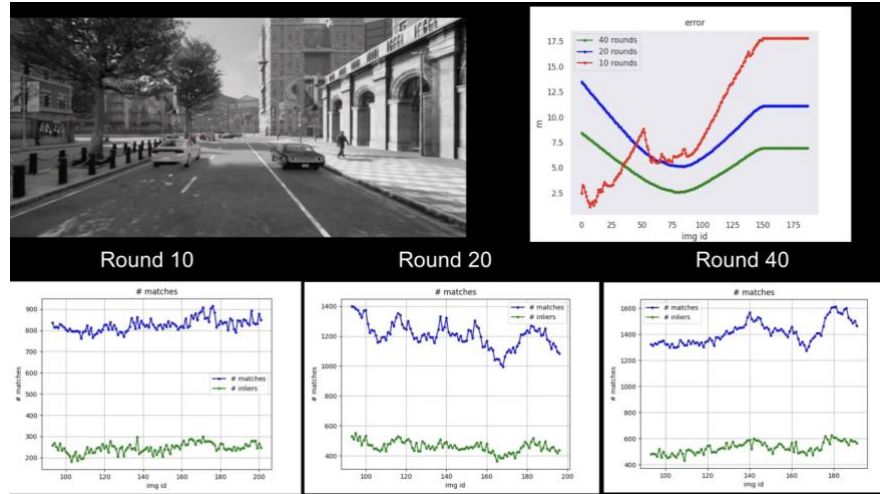


Figure 36. Impact of FL on improving DL-based feature extraction on SLAMs

PoC in the automotive domain: applying HE in FL using the averaging fusion rule

In this section we provide details for the experimentation. In particular we utilized the in-house automotive simulator provided by ATHENA to emulate a scenario involving the collaborative training of an AI model for addressing the LiDAR super-resolution problem, employing the HE-FL paradigm. The used AI models were obtained using the deep unrolling (DU) - based XAI-by-design methodology that ATHENA is developing in the framework of T3.6, as part of the Explainable AI module of AIMaaS. The size of the DU models in this experiment was in the order of 100K parameters. The focus of this experiment is to measure the accuracy loss incurred by performing the FL model aggregation in the encrypted domain, compared to the unencrypted (original) version.

To evaluate the HE-FL framework, we examined a network made up of 10 autonomous vehicles, each functioning as a distinct agent with its own private dataset. In this case, each local dataset was sourced from different locations within the CARLA simulator, thereby representing diverse environmental conditions. We determined the number of communication rounds between the central server and the edge devices to be $T=50$.

At each round of FL, the following steps were executed:

1. **Local Training:** Each device optimizes a local model using its private local dataset.
2. **Encryption:** Once the local training is complete, each client encrypts the weights and biases (the parameters) of their local models
3. **Sending Encrypted Denoiser:** Each client then sends their encrypted denoiser to a central server.
4. **Server Aggregation:** The central server receives encrypted models from all clients. Because of the properties of Homomorphic Encryption, it can perform computations on these encrypted models directly. The server aggregates the encrypted parameters from each client to generate an encrypted global model, using the simple averaging fusion rule.
5. **Broadcasting of encrypted global model:** The encrypted global denoiser is then sent back to all the clients. Each client uses their unique decryption key to decrypt the global model.

The Homomorphic encryption and decryption were performed based on the Tenseal library (Benaissa 2021). In particular, the following two functions were used.

- **Encryption:** `context = ts.context(ts.SCHEME_TYPE.CKKS, poly_modulus_degree=4096)`
`context.global_scale = 2**40`

- **Decryption:** encrypted_total.decrypt()

It is noted, here, that during the adopted encryption process (see Figure 37), the polynomial modulus degree that is determined via the variable poly_modulus_degree directly affects:

- The number of coefficients in plaintext polynomials.
- The size of ciphertext elements.
- The computational performance of the scheme, i.e., larger values lead to higher complexity.
- The security level, i.e., the larger the value, the higher the security level.

```
def homomorphic_encryption_process_B(matrix_models, num_clients):
    # setup TenSEAL context
    context = ts.context(ts.SCHEM_TYPE.CKKS, poly_modulus_degree=4096)
    context.global_scale = 2**40
    decrs = []
    start = int(time.time())
    for i in range(matrix_models[0].shape[0]):
        encrypted_rows = []
        for j in range(num_clients):
            row = matrix_models[j][i]
            row = row[row != 0.0]
            enc_row = ts.ckks_vector(context, row)
            encrypted_rows.append(enc_row)
        # add the encrypted rows
        encrypted_total = encrypted_rows[0]
        for enc_row in encrypted_rows[1:]:
            encrypted_total += enc_row
        # decrypt the encrypted row sum
        decr = np.array(encrypted_total.decrypt()).flatten()
        decrs.append(decr)

    decrs = np.array(decrs, dtype='object')
    end = int(time.time())

    return decrs
```

Figure 37. A Python implementation of the homomorphic encryption process that is adopted

The experiments were performed using the Pytorch platform on an Ubuntu dockerized system with the following hardware characteristics: Intel i7-4790K, Graphic card: Nvidia 3060 11 GB VRAM, RAM:32 GB.

The main conclusion of the experiment is that FL-DU with homomorphic encryption achieves the same convergence behavior as compared to the FL-DU without the encryption part. The source code of the experiment can be found in a github repository⁵. This can be observed in Figure 38 which depicts the loss of the derived global model from the proposed deep unrolling FL scheme with and without homomorphic encryption versus the communication rounds. For benchmarking, the loss achieved by the centralized deep unrolling model, is also depicted. As we can see from Figure 20 the proposed FL-DU method is able to achieve similar performance against the centralized solution. Crucially, the proposed method necessitates only a limited number of communication rounds between the server and local agents, along with a mere five epochs of local training per round, to converge effectively to the centralized solution. Although, the centralized scheme attains marginally superior quantitative results, it necessitates the exchange of considerable amounts of data, thus imposing a considerable burden on the communication links between edge devices and the central server, and raising data privacy concerns. In contrast, the proposed FL-DU scheme provides an efficient solution that overcomes these issues by requiring agents to share only their local denoisers (or priors) from the respective deep unrolling models, which capture detailed information regarding the structure and dependencies of the range images. Another important aspect that stems from the proposed federated unrolling strategy is the fact that we can incorporate any privacy preserving strategy.

⁵ <https://github.com/alexandroskg/Federated-Deep-Unrolling-Lidar-Super-resolution-SLAM>

Interestingly, the FL-DU with the homomorphic encryption achieves the same convergence behavior as compared to the FL-DU without the encryption part.

The use of deep unrolling models in the FL-DU framework provides several benefits. Firstly, these models possess a concise structure, resulting in reduced communication overhead between the server and local agents. This reduction is due to the relatively small number of parameters compared to other state-of-the-art deep learning models. Secondly, the FL-DU framework requires a minimal number of communication rounds to converge to the centralized deep unrolling solution. Notably, it only takes 20 rounds for the FL-DU framework to converge to the centralized DU solution. This efficiency in terms of both communication rounds and model parameters highlights the advantages of the FL-DU framework in federated learning settings.

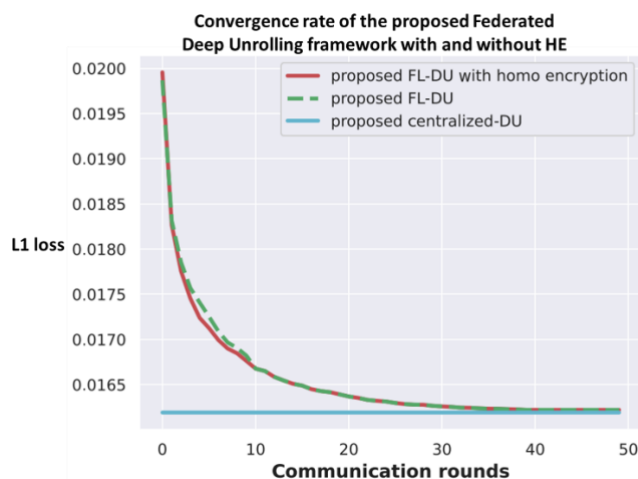


Figure 38. L1 loss of the derived global model from the proposed deep unrolling FL scheme with and without homomorphic encryption vs communication rounds along with the best accuracy achieved by the centralized deep unrolling model.

4.6.1.3. EXPLAINABLE AI

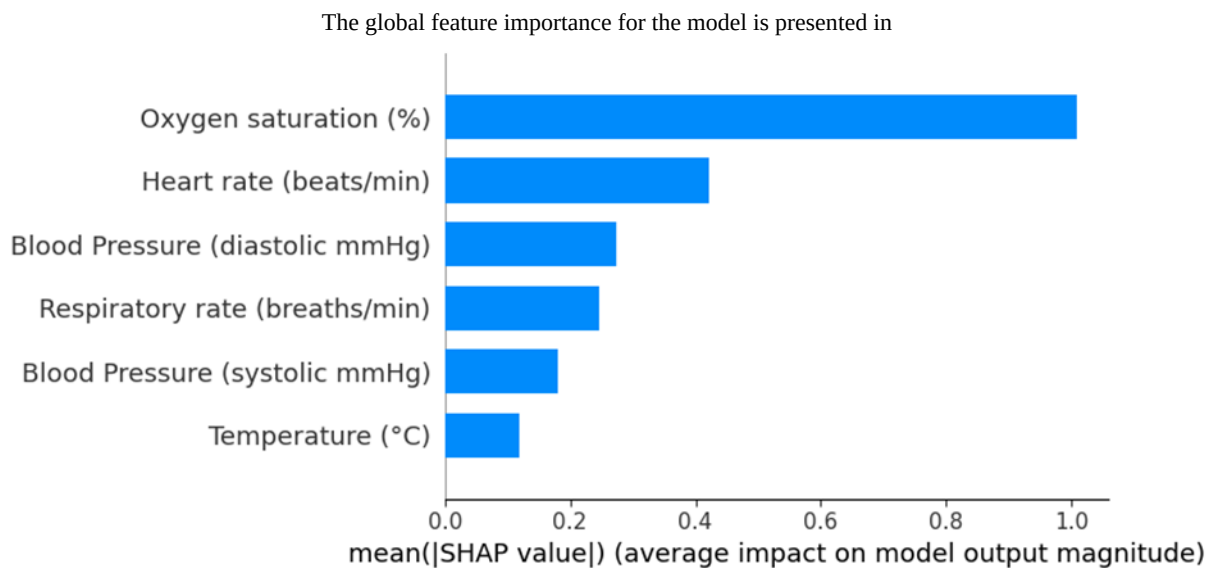
Validation of the xAI subcomponent (Health Pilot)

The XAI subcomponent was thoroughly validated using a dataset from the Health Pilot, encompassing vital measurements. A regressor was trained to predict the Glasgow coma score, utilizing essential features such as temperature, heart rate, respiratory rate, and blood pressure. For model training, Scikit-learn was employed to build linear regression, decision tree, and random forest models, while a dense neural network was trained using Keras/TensorFlow. Following successful model training, the pre-processed data and trained models were exported.

To evaluate the subcomponent, the exported dataset and models were then loaded into the XAI subcomponent. Utilizing the powerful SHAP method, the subcomponent effectively computed feature importances and generated detailed explanations. To gain a deeper understanding and transparency, the subcomponent produced several insightful plots for visualizing the explanations.

Presented below are the results obtained from the XAI subcomponent, utilizing the health dataset and the random forest regressor. These results offer valuable insights into the interpretability of the model's predictions and shed light on the influential features that drive the Glasgow coma score predictions.

The global feature importance for the model is presented in **Error! Reference source not found.** As it can be observed, the feature that has the highest influence on model predictions is “Oxygen saturation (%)”.



. As it can be observed, the feature that has the highest influence on model predictions is “Oxygen saturation (%)”.

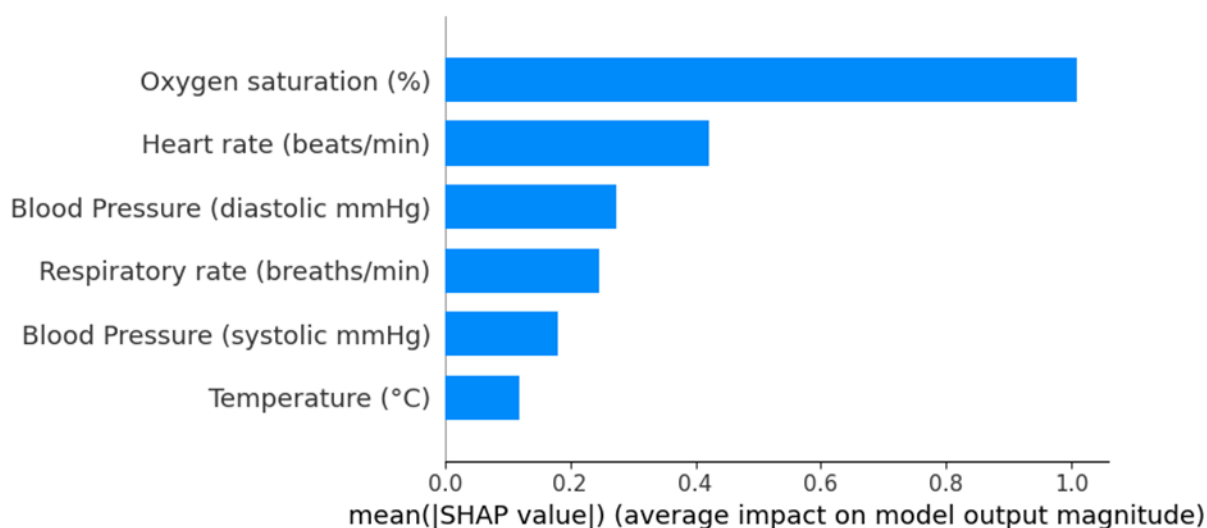


Figure 39: SHAP Feature Importance Plot

The SHAP summary plot (Figure 40) presents additional information, about how the feature values affect the output of the model.

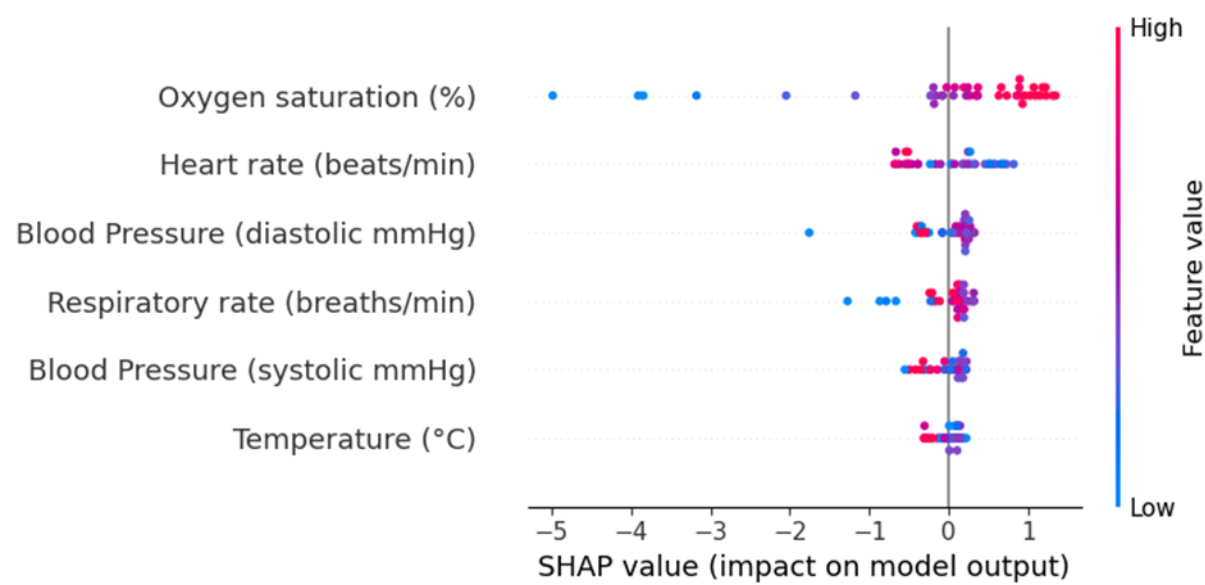


Figure 40: SHAP Summary Plot

The SHAP Dependence plot for one feature (Figure 41) shows the SHAP values for the different values of “Temperature” feature.

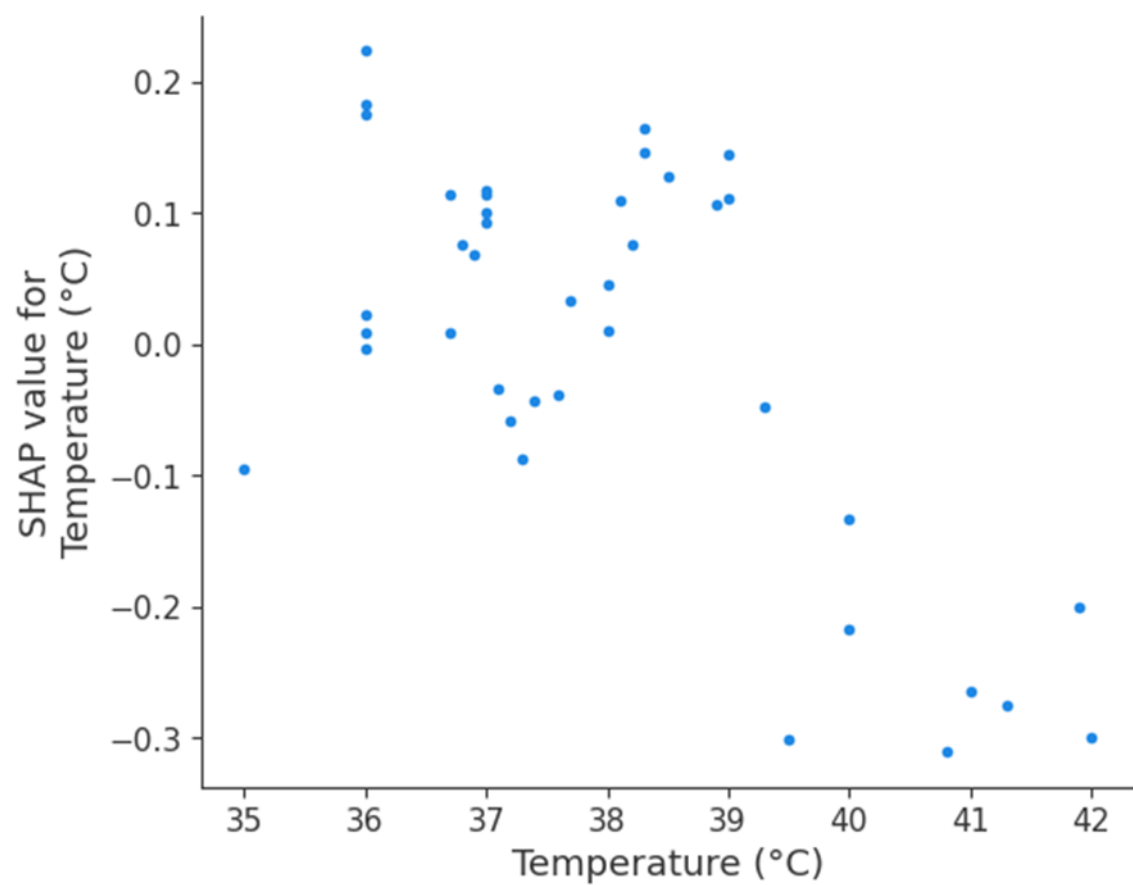


Figure 41: SHAP Dependence Plot

The SHAP Dependence with Interaction Plot (Figure 42) presents how two features, “Temperature” and “Oxygen saturation”, interact in terms of SHAP values.

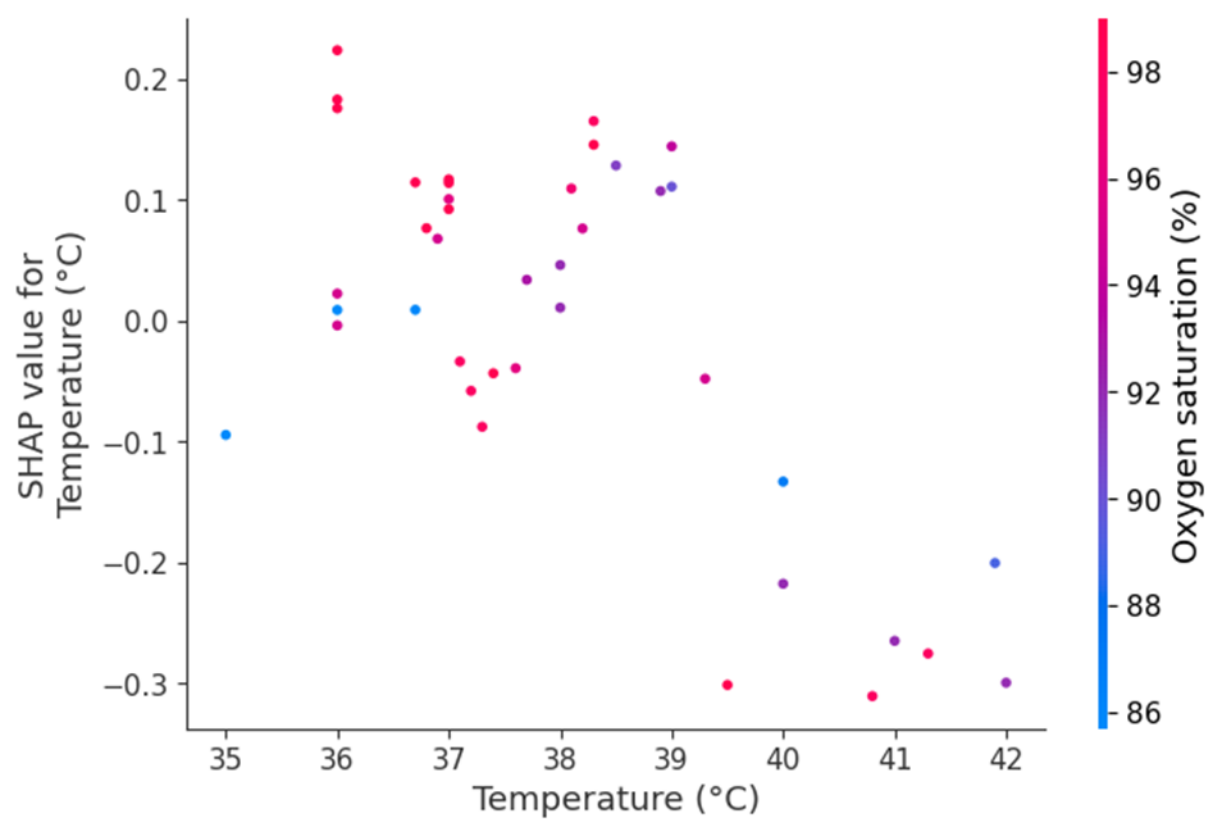


Figure 42: SHAP Dependence with Interaction Plot

The SHAP Force plot (Figure 43) explains a single instance by visualizing the SHAP values as “forces”.



Figure 43: SHAP Force Plot

In Figure 44 the SHAP force plots for all the instances are clustered by similarity, providing another form of model global explanation.

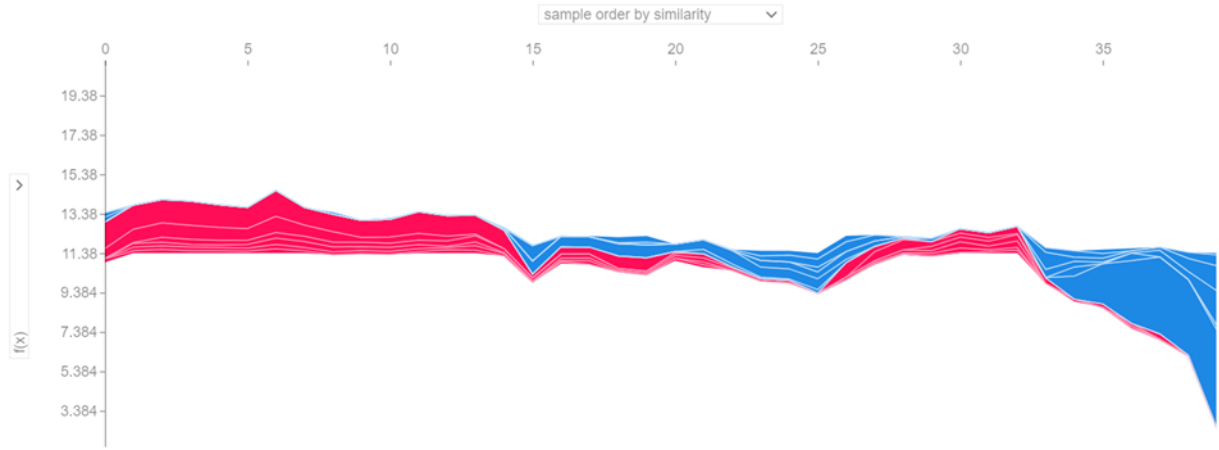


Figure 44: SHAP explanations clustered by similarity Plot

The SHAP Waterfall plot (Figure 45) explains a single instance, illustrating how the model reaches the final prediction from the baseline value. In this case, the feature “Respiratory rate” has the biggest impact on model prediction.

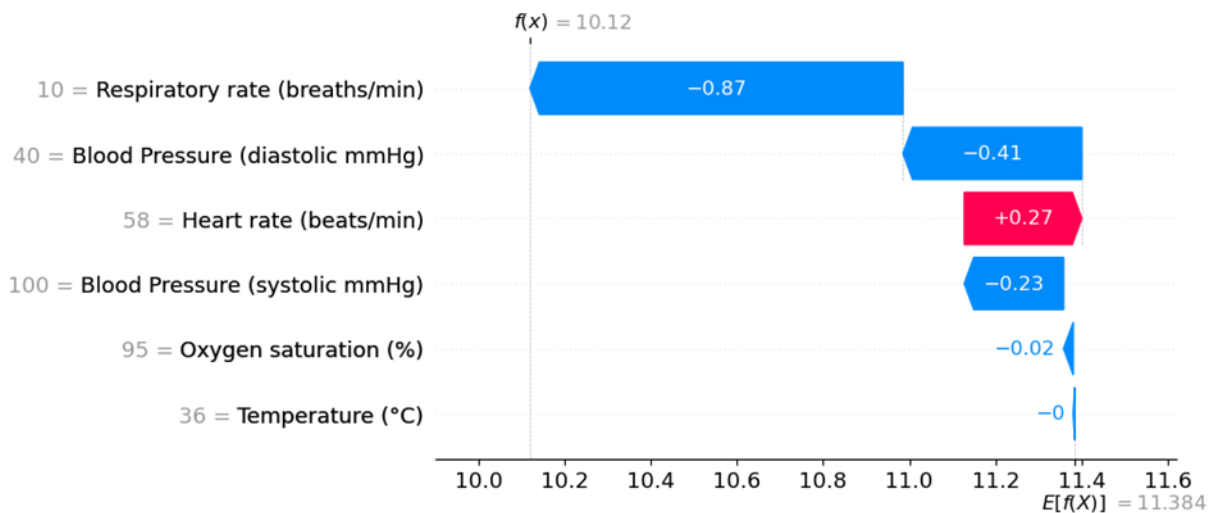


Figure 45: SHAP Waterfall Plot

The comprehensive validation process of the XAI subcomponent ensures that it is a reliable and powerful tool, capable of providing meaningful insights into complex machine learning models. With its visualization capabilities and precise explanations, the XAI subcomponent empowers users to make informed decisions, enhancing the interpretability and reliability of the model's predictions.

Validation of the XAI-by-design (Automotive)

In the following, an illustrative example is provided to make this methodology more comprehensive by focusing on the problem of LiDAR super resolution in the automotive domain. Consider a high-resolution point cloud derived from a 64-channel LiDAR sensor and the respective high-resolution range image $X \in \mathbb{R}^{C \times M}$, where C represents the vertical resolution (i.e., the number of channels or lasers such as $C = 64$) and M denotes the horizontal resolution of the range image. Mathematically, the corresponding low-resolution range image $Y \in \mathbb{R}^{c \times m}$ with the same

horizontal resolution as $\mathbf{X}$ but only $c < C$ channels in the vertical resolution (e.g., $c = 16$) can be derived based on the following degradation model (Shan 2020).

$$\mathbf{Y} = \mathbf{S}\mathbf{X} + \mathbf{N},$$

where $\mathbf{S} \in \mathbb{R}^{c \times C}$ denotes the downsampler operator that selects only the c channels from the high resolution range image and $\mathbf{N}$ is a zero-mean Gaussian noise term. Given the low-resolution range image $\mathbf{Y}$, our goal is to effectively estimate the corresponding high resolution range image $\mathbf{X}$. To achieve this goal, the above minimization problem can be adopted. Note that the high-resolution point cloud is derived by transforming the estimated high-resolution range image into 3D coordinates (see Figure 46. Given a low-resolution 3D point cloud derived from a 16-channel lidar, it is project onto a 2D low-resolution range image. This image is provided as input to the proposed deep unrolling model to estimate the corresponding high-resolution range image. The estimated image is then transformed into 3D coordinates, thus producing the desired high resolution point cloud for the lidar odometry problem.).

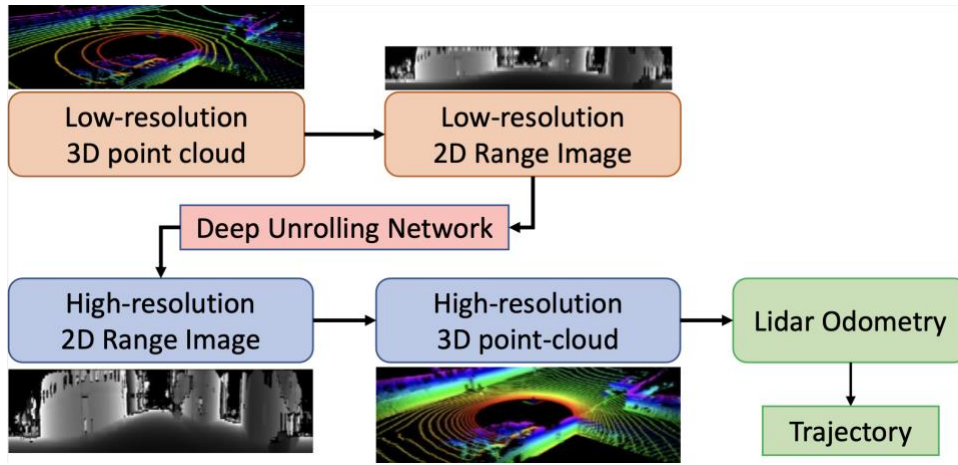


Figure 46. Given a low-resolution 3D point cloud derived from a 16-channel lidar, it is project onto a 2D low-resolution range image. This image is provided as input to the proposed deep unrolling model to estimate the corresponding high-resolution range image. The estimated image is then transformed into 3D coordinates, thus producing the desired high resolution point cloud for the lidar odometry problem.

Solving the problem using the Iterative solver HQS

According to the first step of the methodology, the HQS algorithm is employed to solve the minimization problem under consideration (Geman 1995). By introducing an auxiliary variable $\mathbf{Z} \in \mathbb{R}^{C \times M}$, the problem can be written as a follows

$$\underset{\mathbf{X}}{\operatorname{argmin}} \left(\frac{1}{2} \|\mathbf{Y} - \mathbf{S}\mathbf{X}\|_F^2 + \mu R(\mathbf{Z}) \right), s. t. \mathbf{Z} - \mathbf{X} = 0,$$

The loss function that HQS aims to minimize is

$$\mathcal{L} = \frac{1}{2} \|\mathbf{Y} - \mathbf{S}\mathbf{X}\|_F^2 + \mu R(\mathbf{Z}) + \frac{b}{2} \|\mathbf{Z} - \mathbf{X}\|_F^2 \quad (1)$$

where b is a penalty parameter. Based on (1), a sequence of individual sub-problems emerges:

$$\mathbf{X}^{k+1} = \underset{\mathbf{X}}{\operatorname{argmin}} \left(\frac{1}{2} \|\mathbf{Y} - \mathbf{S}\mathbf{X}\|_F^2 + \frac{b}{2} \|\mathbf{Z}^k - \mathbf{X}\|_F^2 \right), \quad (2)$$

$$\mathbf{Z}^{k+1} = \underset{\mathbf{Z}}{\operatorname{argmin}} \left(\mu R(\mathbf{Z}) + \frac{b}{2} \|\mathbf{Z} - \mathbf{X}^{k+1}\|_F^2 \right). \quad (3)$$

The sub-problem in (2) constitutes a quadratic regularized least squares problem. The closed form solution is given as

$$\mathbf{X}^{k+1} = (\mathbf{S}^T \mathbf{S} + b\mathbf{I})^{-1} (\mathbf{S}^T \mathbf{Y} + b\mathbf{Z}^k). \quad (4)$$

Moreover, from a Bayesian perspective, sub-problem (3) corresponds to a Gaussian denoiser (Zhang 2017). Hence, a neural network $f_\theta(\cdot)$ can be employed to act as denoiser and learned from training data. Thus, (3) can be written as

$$\mathbf{Z}^{k+1} = f_\theta(\mathbf{X}^{k+1}). \quad (5)$$

Thus, the HQS solver consists of the update rules described by (4) and (5). Note that the neural-network $f_\theta(\cdot)$ can be pre-trained employing pairs of generated noisy high resolution range images (using Gaussian noise, denoted as $\mathbf{N}$) and their corresponding ground truth versions denoted as $\{\mathbf{X}^i + \mathbf{N}, \mathbf{X}^i\}_{i=1}^p$ via the loss function $\sum_{i=1}^p \|\mathbf{f}_\theta(\mathbf{X}^i + \mathbf{N}; \theta) - \mathbf{X}^i\|_F^2$.

Interpretable deep unrolling-based model

In this section, the deep unrolling framework is applied. In particular, a small number of iterations, say K , of the proposed HQS solver are unrolled and treated as a deep learning architecture. Each iteration of the iterative solver is considered a unique layer of the proposed model, resulting in a K -layer deep learning architecture. In light of this, the parameters and the depth of the derived model are highly interpretable as they have direct correspondence to the underlying iterative algorithm. Figure 47 demonstrates the proposed deep unrolling network.

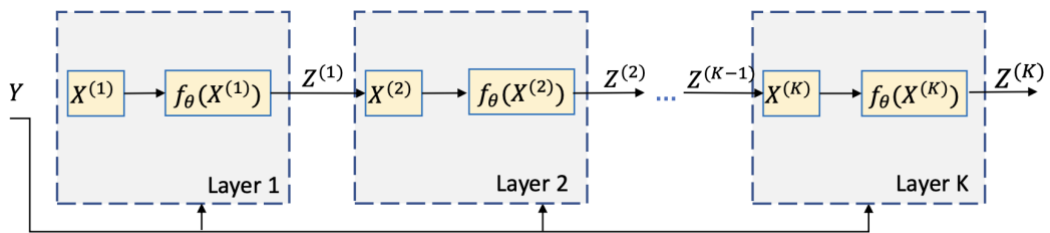


Figure 47. The proposed end-to-end Deep unrolling model.

A great benefit that stems from this strategy is the fact that the learnable parameters of the proposed model (i.e., the neural network $f_\theta(\cdot)$ and the penalty parameter b) can be trained end-to-end by minimizing some loss function, as follows

$$l(\theta) = \sum_{i=1}^p \|z_i^{(K)} - x_i\|_F^2, \quad (6)$$

Where $\mathbf{Z}_i^K$ is the output of the proposed deep unrolling network given a low-resolution range image $\mathbf{Y}$ and $\mathbf{X}$ denotes the i -th ground truth high resolution range image.

Validation on super-resolution LiDAR data

The functionality of the XAI-by design module has been validated for the problem of super-resolution LiDAR data, namely, acquiring costly 64-channel LiDAR data from more affordable, though of lower-resolution, 16-channel LiDAR data. The super-resolution DNN is designed and trained using the Deep Unrolling (DU) methodology as devised in the frame of T3.6 and the XAI-by design module. The resulting DNN is then used in a sequence of frames capturing a traffic scenario in an urban environment and provides the input to a LiDAR SLAM approach. The results are depicted in Figure 48. At the left-hand column, a snapshot of the traffic conditions is depicted at the upper image while the data gathered by the 16-channel LiDAR sensor are presented at the bottom. In the other two columns (bottom images), the 64-channel reconstructed (from the 16-channel case) and ground-truth LiDAR data are presented, respectively, to validate the high quality of reconstruction. Moreover, at the upper part of those columns, the improvement on the localization error of using the reconstructed data versus the low-resolution data in the LiDAR SLAM approach, are observed.

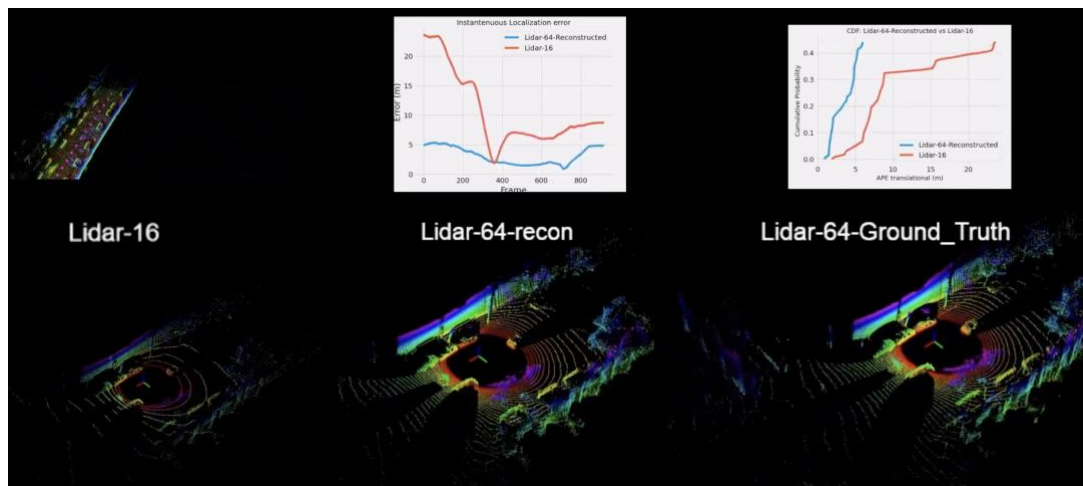


Figure 48. Impact of Explainable AI DU-model on improving raw data quality and the benefits on LiDAR SLAM

4.7. NEXT STEPS

AIMaaS

This module is going to be fully utilized and validated within the platform. The goal is to generalize and validate its use through different scenarios and domains. This will be an ongoing process – that will allow at the end to be fully general supporting dynamic selection criteria for modes in order to support AI through TRUSTEE as a service. Our goal is to fully integrate AIaaS with in the TRUSTEE Platform.

Federated Learning & xAI-by-design

The next steps regarding the AIMaaS/FL module will focus on implementing the functionality pertaining to performing multiple rounds of FL (and the required FL orchestration), including a) tracking FL rounds and orchestrate the collection of local models for the aggregation step by the HEDF; b) Storing the Global Model after k rounds to the DL, and sharing it with the users that participate in the FL group. Additionally, more advanced FL fusion rules such as weighted and extended aggregation are going to be investigated. Regarding the latter, which occurs as a synergy between the AIMaaS/FL and explainable AI modules, we will focus on implementing the required functionalities for initializing a deep unrolling model design, which entails a) selection of regularizer model architecture; b) selection of employed solver; c) selection of number of iterations (network depth).

Explainable AI

In the context of Task 3.6 "Explainable AI functions", the xAI subcomponent of the AIMaaS ABB has been developed. This tool will be incorporated in the AI model as a service and will provide comprehensive explanations of the involved AI models. Leveraging various state-of-the-art explainability techniques, the xAI subcomponent aims to enhance the transparency and interpretability of AI models, ensuring that users can gain insights into the reasoning behind their decisions. This initiative represents a critical step forward in advancing AI transparency and accountability within the platform.

4.8. CONCLUSIONS

This chapter provided an overview of the AIMaaS component and its components. This ABB consists of 3 main sub-components. As explained in the previous sub-chapters, the focus during this first implementation phase of the subcomponents and their internal modules.

5. PART III: TRUSTWORTHINESS AI SUPPORT DESIGN FRAMEWORK (TAI-SDF)

5.1. INTRODUCTION

5.1.1. NEED AND OBJECTIVES

The purpose of this chapter is to define functional and technical components, specifying user requirements, determining system specifications, outlining the overall architecture, and marking out the legal requirements of tools developed in the scope of T3.3. Within T3.3 “Certification and Auditing”, two different tools are developed: the former one is named **Trustworthiness Artificial Intelligence Support Design Framework (TAI-SDF)** and the latter one is named **Accountable Transactions Recorder (ATR)**.

The needs are to ensure:

- **Interoperability:** refers to the capability that enables different systems, devices, applications to seamlessly communicate with each other. The ATR will interoperate with other ABBs enabling seamless integration and data exchange among authorized entities.
- **Transparency:** the immutability nature of blockchain will ensure tamper-resistant transactions; thus, enabling transparency in the recorded transactions.
- **Traceability:** Blockchain comprises traceability of the recorded transactions; thus, enables to achieve a transparent audit trail in order to track all the transaction history.
- **Security:** The blockchain implements cryptographic techniques where every transaction is linked to each other forming blocks. The blocks of transactions are then linked to the adjacent blocks. This innovative process protects us from unauthorized access and also ensures authenticity of the recorded transactions.
- **Trust and Efficiency:** The permissioned blockchain ensures that only the trusted authorities get access to the recorded transactions. Furthermore, the non-requirement of intermediaries also enhances efficiency.
- **Trustworthiness:** includes the following aspects fairness, explainability, security and privacy, robustness of the AI systems through the AI lifecycle; AI system developers and product managers need of a support decision-making tool in order to develop explainable and trustworthy AI applications.

This section will start by providing background about trustworthiness- by-design frameworks and security metrics for AI models. On the second part, the high-level architecture of TAI-SDF will be explained and then move into the description of the detailed architecture.

5.1.2. INNOVATION

The novelty of the TAI-SDF is to enable an easy and guided evaluation of Trustworthiness aspects during the AI system lifecycle. The current technical tools present in the market aim to address specific AI-related issues for some of the Trustworthiness dimensions, namely fairness, transparency, explainability and robustness. Conversely, TAI-SDF is a proprietary methodologically framework enabling auditing, continuously monitoring of AI/ML models and data in compliance with latest AI regulations issued by European Commission and ensuring AI models are fair, robust, reliable, and secure.

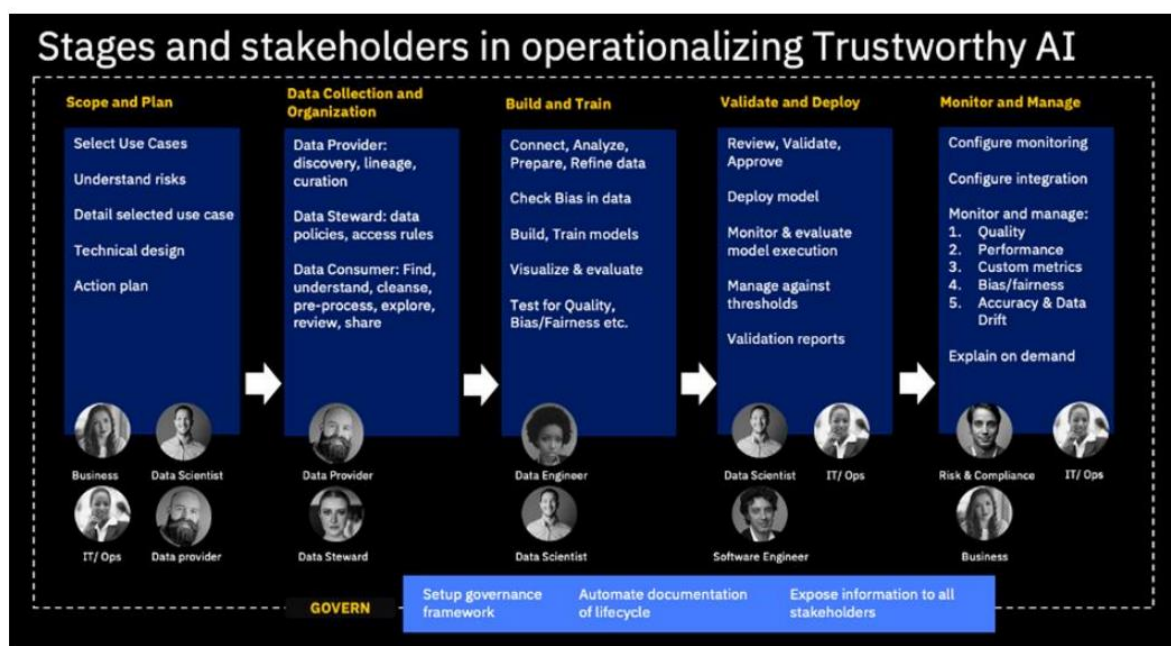
5.2. BACKGROUND

5.2.1. TRUSTWORTHINESS-BY-DESIGN FRAMEWORKS FOR MONITORING, AUDITING AND CONFORMITY CHECK

AI systems are playing an important role as part of decision and control systems in economic and social sectors. Given the superior performance of AI technologies as compared to traditional IT technologies, technical, academic, and business communities can benefit from AI applications by reducing labour and promoting social good. However, recent research indicate that AI can cause unintentional harm to humans by making unreliable decisions in safety-critical scenarios or undermining fairness by inadvertently discriminating against a group or groups (Liu 2022). Consequently, a large body of literature on trustworthy AI has emerged; trustworthy AI can be defined as programs and systems built to solve problems like a human, which bring benefits and convenience to people with no threat or risk of harm. A systematic approach is needed to build new AI applications in a trustworthy manner; this approach must take into consideration the end-to-end data science and AI lifecycle. The AI lifecycle consists of a sequence of stages that bring together various personas, such as Data Scientist, AI Engineer, Business owner and Product Manager. The stages of the lifecycle shall include (Li 2022):

- Scope and plan: this stage guides the prioritization of the use case scenarios and development of an AI action plan leading to business requirements definition.
- Data Management: during this stage data are collected and pre-processed, for instance features selection.
- Algorithm Design and Development: during this stage data science teams explore and prepare the data, build, train, and test their AI/ML models. Bias should be checked before building the model. Model robustness, explainability and other aspects can be tested. Once these tests complete successfully, an MLOps pipeline allows for the model and related assets to be moved from the development environment to a pre-production validation environment.
- Verification, Validation and Deployment: this stage involves validation of the model and deployment into production. Validation activities could be performed by a team other than the one that built the model, such as the organization's model validation or model risk management team or an external entity. This team validates quality, robustness, fairness, etc. and generates validation reports.
- Operation: in this stage AI/ML models are monitored. Quality, robustness, and fairness are periodically monitored according to the business need.

Considering that AI Trustworthiness needs to be constantly monitored along all AI stages, a new industrial AI workflow has been introduced. Recent studies extended the DevOps paradigm to MLOps to cover the ML lifecycle of data preparation, training, validation, and deployment.

Figure 49: AI system lifecycle vs actors per each stage ⁶

5.2.2. SECURITY METRICS FOR AI MODELS

The dependency of libraries ecosystems and the presence of vulnerabilities in such dependencies have been studied in the literature (Decan 2019, ^{20,21}. In (Decan 2018), the authors studied the *npm* dependency packages. They found that the number of packages with a known security vulnerability is growing over time. In (Imtiaz 2021) the authors compared the analysis reports of software composition analysis tools on a large web application, OpenMRS, composed of Maven (Java) and *npm* (JavaScript) projects. In (Nicolae 2018), the authors proposed Adversarial Robustness Toolbox (ART) for Adversarial robustness of AI. Adversarial robustness for AI refers to the ability of an AI system to resist and defend against adversarial attacks. Adversarial attacks are deliberate attempts to manipulate an AI system by introducing small, often imperceptible, changes to the input data that can cause the system to make incorrect predictions or decisions. Adversarial robustness is important for ensuring the reliability and security of AI systems, particularly in applications such as autonomous vehicles, medical diagnosis, and financial forecasting. Techniques for improving adversarial robustness include adversarial training, input pre-processing, and model regularization. In (Mahmoud 2023), the authors presented an empirical study of 1,396 vulnerability reports affecting 698 Python packages in the Python ecosystem (PyPi). Their findings show that the discovered vulnerabilities in Python packages are increasing over time, and they take more than 3 years to be discovered. These vulnerabilities (40.86%) are only fixed after being publicly announced, giving a time for attackers to exploit the vulnerability.

Recently, several works have been proposed in adversarial attacks and defenses in various domains of AI models. Zhang et al (Zhang 2020) provide an overview of adversarial attacks and defenses in deep learning, covering various attack methods and defense strategies. Wang et al (Wang 2021) survey adversarial attacks and defenses in natural language processing, including attacks on text classification, sentiment analysis, and machine translation. Chen et al (Chen 2022) review adversarial attacks and defenses in graph neural networks, covering attacks on node

⁶ Foundations of trustworthy AI: Operationalizing trustworthy AI <https://www.ibm.com/blog/operationalizing-trustworthy-ai/>

classification, link prediction, and graph classification. Liu et al (Liu 2023) provide a survey of adversarial attacks and defenses in computer vision, including attacks on image classification, object detection, and semantic segmentation. Wang et al (Wang 2020) survey adversarial attacks and defenses in reinforcement learning, covering attacks on policy gradients, Q-learning, and actor-critic methods. Zhang et al (Zhang 2021) review adversarial attacks and defenses in federated learning, covering attacks on model aggregation, client selection, and communication. Chen et al (Chen 2021) survey adversarial attacks and defenses in time series classification, covering attacks on various types of time series data, such as sensor data, medical data, and financial data. These works provide a comprehensive overview of the current state-of-the-art in adversarial attacks and defenses in various domains of AI models. It highlights the importance of developing robust and secure AI systems. The need of security metrics for AI systems to defend against adversarial attacks due to the difficulty of achieving adversarial robustness, which requires techniques such as adversarial training, input pre-processing, and model regularization. Additionally, there is a risk that AI and ML could be used to create new types of attacks that are not yet fully understood or anticipated. Therefore, developing effective measures against AI and ML-based attacks is a critical and challenging task.

5.3. REQUIREMENTS

In this chapter, the compliance of the TAI-SDF ABB of TRUSTEE is investigated with regards to the user-centred, technical, legal, and socio-ethical requirements that were reported in D2.1 (TRUSTEE consortium, 2023). The results of this investigation provide further preliminary insight regarding with the aforementioned requirements addressed during the development and by the implementation of the TAI-SDF and are presented in the following sub-sections. It is worth noting that only the Mandatory, Desirable, and Optional requirements have been considered at this point. Furthermore, only the currently compliant or partially-compliant requirements have been considered. As the implementation of the DPIA ABB progresses during the development of the TRUSTEE and the maturity of the ABB increases, the investigation of the compliance will evolve and further updates and/or enhancements will be provided in upcoming versions of this deliverable, namely D3.2 and D3.3.

User-centred requirements were identified through the UX Research that was conducted under T2.3. These requirements were reported in D2.1 (TRUSTEE consortium, 2023), while they were also utilized in order to extract technical, functional and non-functional requirements, that were also outlined in D2.1. Table 12 below presents the interplay between user-centred requirements and technical requirements, that is the functional and non-functional requirements that address each user-centred requirement. The user-centred requirement ID is used to reference the respective requirement reported in D2.1 (TRUSTEE consortium, 2023) while the requirement's description is also provided alongside whether the DPIA ABB is currently fully or partially compliant with the user-centred requirement, i.e., indicating whether technical requirements have been identified in order to address the respective user-centred requirement. The ID of the technical requirement, as reported in D2.1 (TRUSTEE consortium, 2023), that is addressing each user-centred requirement is also provided in the last column of the table. Additionally, it is possible than more than one technical requirement addresses a single user-centred requirement.

Table 12: TAI-SDF – User-centred requirements

Req id (WP2)	Req. Description	Compliance	How is addressed
USR-014	TRUSTEE should offer APIs or portals for some of its functionalities to be used.	Y	Req-TAI-SDF-NFUNC-1

USR-037	TRUSTEE should be able to assess the compliance of an AI model to policies, regulations, and standards.	Y	Req-TAI-SDF-FUNC-1, Req-TAI-SDF-FUNC-2, Req-TAI-SDF-FUNC-3
USR-091	TRUSTEE could assess and monitor the performance of the AI models regarding fairness, explainability, robustness, and security.	Y	Req-TAI-SDF-FUNC-1, Req-TAI-SDF-FUNC-2, Req-TAI-SDF-FUNC-3
USR-093	TRUSTEE could support fully trustworthy AI/ML models.	Y	Req-TAI-SDF-FUNC-1, Req-TAI-SDF-FUNC-2, Req-TAI-SDF-FUNC-3

Table 13 presents the preliminary results of the investigation regarding the compliance of the TAI-SDF with the technical, functional, and non-functional requirements specified in D2.1. The ID of the respective technical requirement is used as a reference to D2.1, while the requirement's description is also provided alongside whether DA complies with the requirement and how. Further updates and/or enhancements regarding the approach presented below for addressing each technical requirement will be provided in upcoming versions of this deliverable, namely D3.2 and D3.3.

Table 13: TAI-SDF – Functional and non-functional technical requirements

Req id (WP2)	Req. Description	Compliance	How is addressed
Req-TAI-SDF-FUNC-1	The TAI-SDF MUST allow the relative TRUSTEE users/personas to define a matrix for phases and principles, assign, modify, delete such phases and principles following the methodology that the framework will implement, in order to ensure the trustworthiness of the AI models	Y	TAI-SDF will allow to edit a new phase of AI lifecycle and delete a phase of the AI lifecycle. TAI-SDF will allow to edit a new principle so that the TRUSTEE users can evaluate it and to delete a principle so that he/she can delete irrelevant components.
Req-TAI-SDF-FUNC-2	The TAI-SDF MUST support the creation, modification, deletion of relevant templates and questionnaires regarding the validation of trustworthiness that of the AI models and associate them with the phase and principles matrix.	Y	TAI-SDF will allow to create multiple templates, to enter scale and metrics so that each question can be quantitatively evaluated, to delete questionnaires so that User can delete unnecessary questionnaires.
Req-TAI-SDF-FUNC-3	The TAI-SDF MUST communicate to the TRUSTEE users/personas that a project related to the AI model under development has been created and allow access to information such as: the creation date, last change saved, which is the phase and principle currently under evaluation in the project, so that the TRUSTEE user/persona can select to start or continue an assessment and visualise the assessment output and the received score at each stage along with recommendations		TAI-SDF will allow to manage projects and create new ones. TAI-SDF will allow to start the assessment of the chosen phase following the AI lifecycle model and to complete the checklist/questionnaire.

Req-TAI-SDF-NFUNC-1	provided by the tool for improving the score. The TAI-SDF SHOULD be a web application.	Y	TAI-SDF will allow to view a webpage with menu items for accessing services, notifications list, language choice.

A preliminary analysis of the compliance with the legal and socio-ethical requirements defined in the D2.1 (TRUSTEE consortium, 2023) is shown in ANNEX 2 and ANNEX 3.

5.4. ARCHITECTURE

5.4.1. HIGH LEVEL

The overall high-level architecture of the TAI-SDF is depicted in the following diagram; the TAI-SDF architecture consists of several subcomponents Administration module, Authentication module, HomePageModule, OperationModule, SurveyModule and GDPR Compliance.

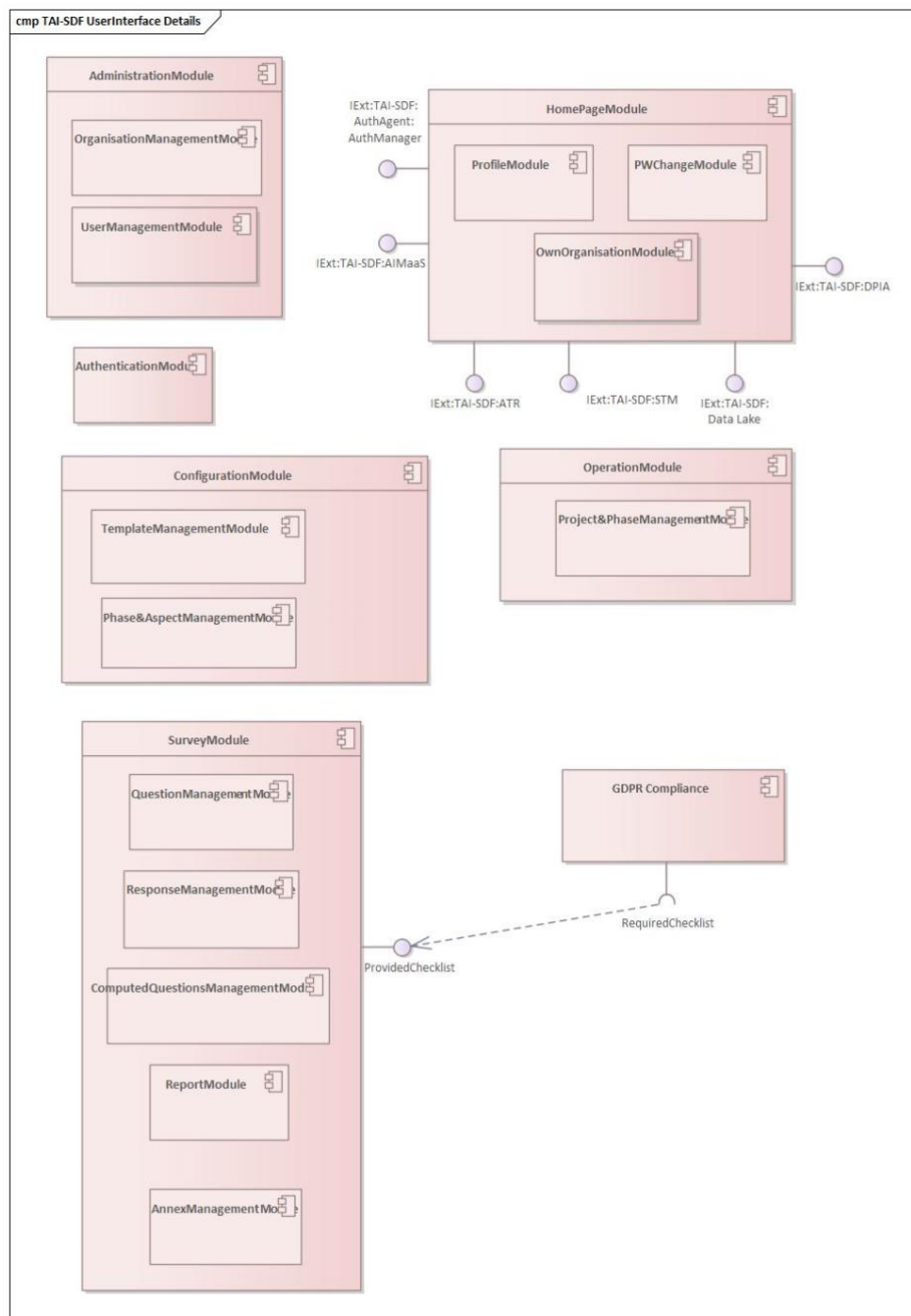


Figure 50: TAI-SDF overall architecture

5.5. IMPLEMENTATION

5.5.1. FUNCTIONAL DESCRIPTION

As mentioned earlier, the TAI-SDF architecture is composed by the following subcomponents:

- **ConfigurationModule** that contains all the functions and procedures to define matrixes and templates;

- **HomePageModule** that contains all the functions and procedures to access to organization and profile management capabilities and to the external interface of the TAI-SDF;
- **OperationalModule** that contains all the functions and procedures to create, modify and execute projects;
- **AdministrationModule** that contains all the functions, and procedure to manage organization and users;
- **SurveyModule** that contains all the functions, and procedure to define and execute questionnaires and tests;
- **GDPR Compliance** that contains all the functions and procedures to manage the GDPR checklist;
- **AuthenticationModule** that contains all the functions and procedures to interface with the authentication agent.

5.5.2. OPERATIONAL DESCRIPTION

The users that will be able to interact with the TAI-SDF tool include the **TRUSTEE user** and **TAI-SDF user**:

- The **TRUSTEE user** (Model/Data Provider, Consumer and Developer) will be able to launch the TAI-SDF tool directly from AIMaaS in order to evaluate trustworthiness aspects; moreover, the TRUSTEE user will be able to launch the GDPR compliance checker, an inner module of TAI-SDF, from STM in order to perform the GDPR assessment. The user will be able to assess:
 - The aspects concerning trustworthiness are considered in the questionnaire which is envisioned to be a part of the TAI-SDF tool and presented in ANNEX 4: TAI-SDF Questionnaire for the assessment of the compliance with trustworthiness requirements .
 - The aspects concerning GDPR compliance in the context of the data provision, access, and exchange of Data Provider and (Data) Consumer are considered in the questionnaire included in ANNEX 5: GDPR Compliance Checklist for TAI-SDF and STM.
- The **TAI-SDF user** is the manager who will be able to create the project, associates it with a template not modifiable, create matrix of phases and aspects and assign TRUSTEE user with the project.

Both questionnaires will be further defined and adapted in the whole lifecycle of the project, to enable development of the solution responding to the needs of the users in the best possible manner. Further modifications of these initial questionnaires will be reflected in D3.2 and D3.3. The sequence diagram of the TRUSTEE user flow shows the steps that will be performed within the TRUSTEE platform, in the case a model provider logged on the AIMaaS wants to evaluate the trustworthiness of the AI model he would like to upload or in case a TRUSTEE user willing to run explainability functions on AIMaaS will be redirected to TAI-SDF in order to perform self-assessment and get the trustworthiness report. Following Figure 51, the TRUSTEE user will access to the Dashboard, then they will be redirected to the AIMaaS module. Then, the AIMaaS will check the authentication of the TRUSTEE user through the Authentication Manager. Afterwards, the TRUSTEE user will be redirected to the TAI-SDF in order to evaluate the Trustworthiness principles. Figure 52 shows the sequence diagram of how TAI-SDF user is authenticated: the manager needs to be registered and authenticated in order to create the project, the matrix of the phases and assign TRUSTEE user with the project. Another scenario is depicted in Figure 54 where the Data Provider needs to subscribe an agreement whose validity is checked by the GDPR Compliance checker. The Data Provider and Data Consumer will log in the Security and Trust Manager, then they will be redirected to the GDPR Compliance checker where to perform the self-assessment.

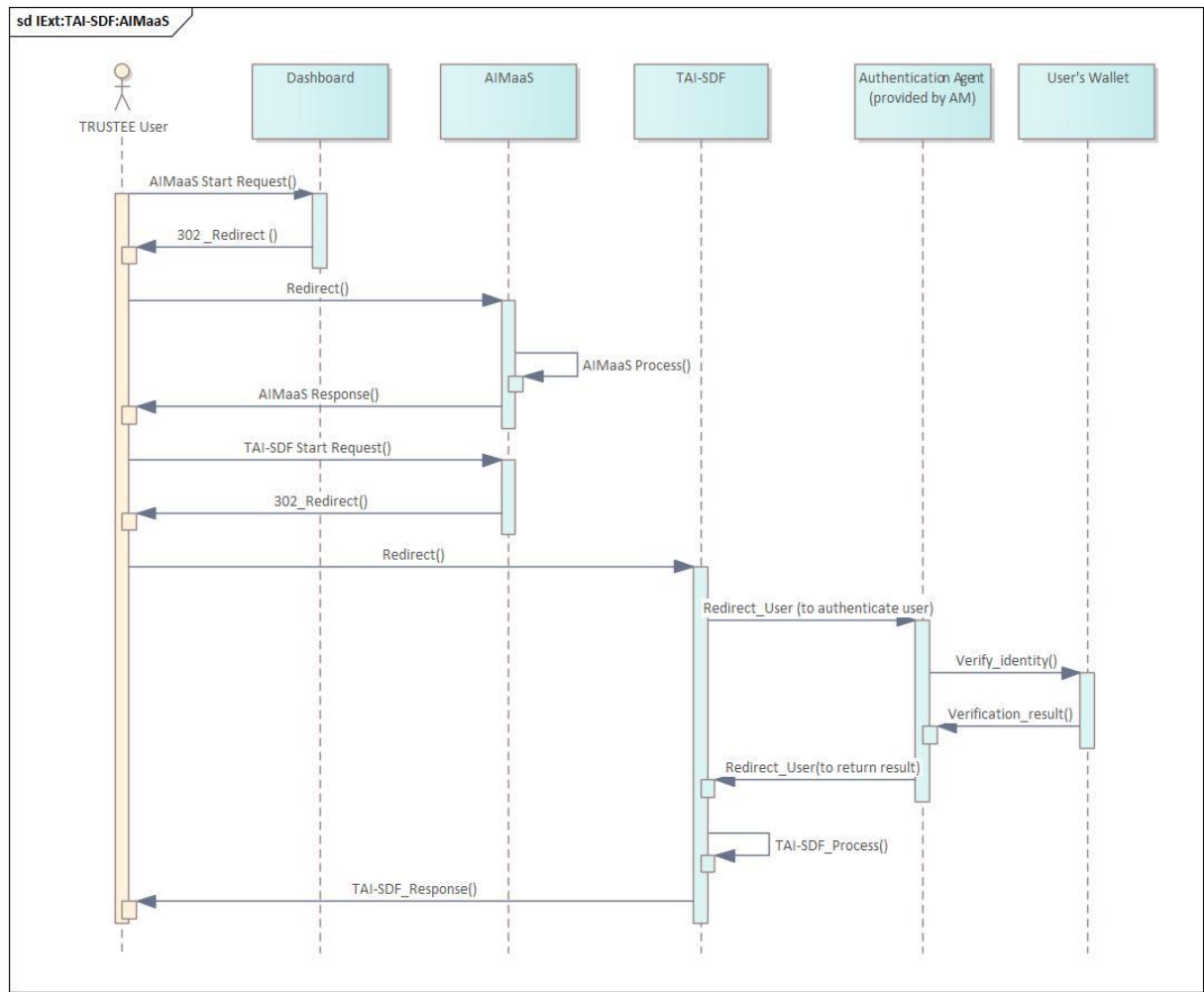


Figure 51: Sequence diagram related to the interaction between TAI-SDF and AIMaaS

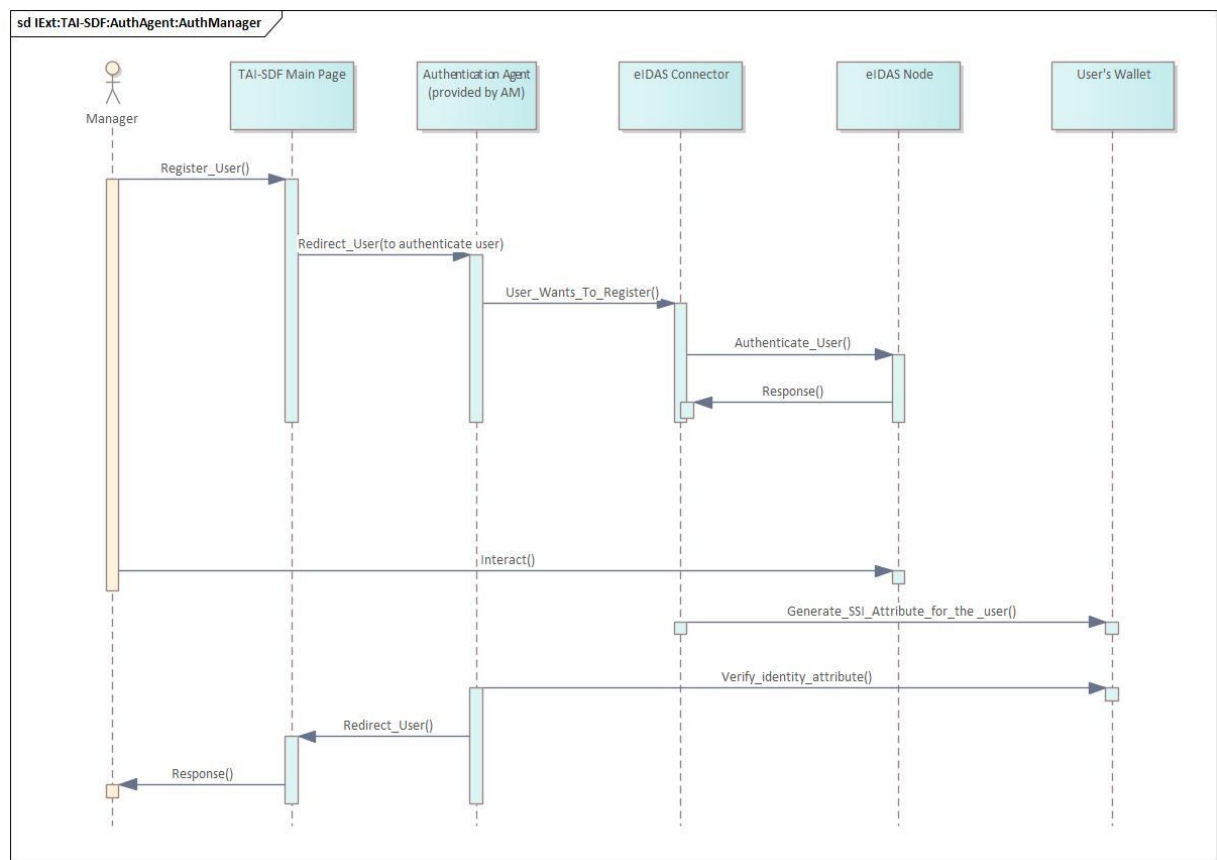


Figure 52: Sequence diagram showing the process of TAI-SDF user authentication

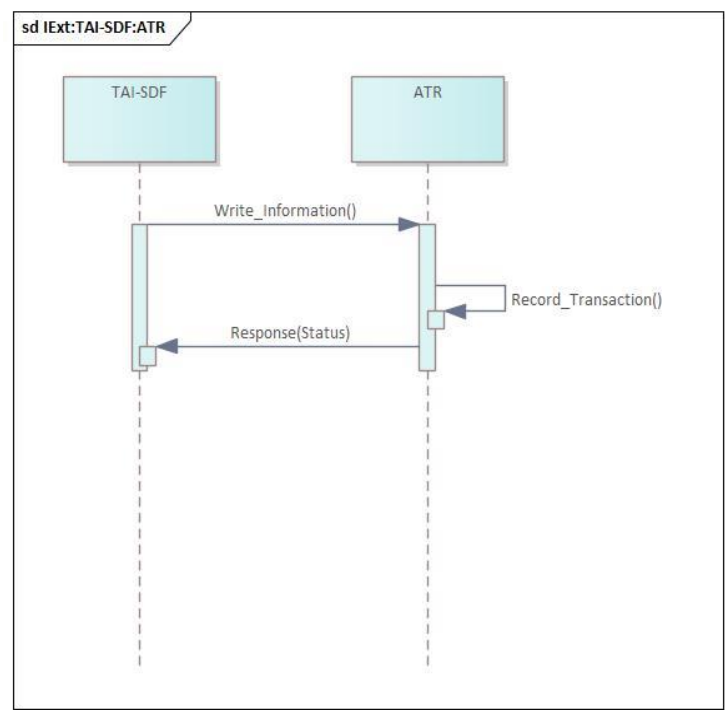


Figure 53: Sequence diagram showing the process of recording activities on ATR , such as add answers to questions, modify answers to questions

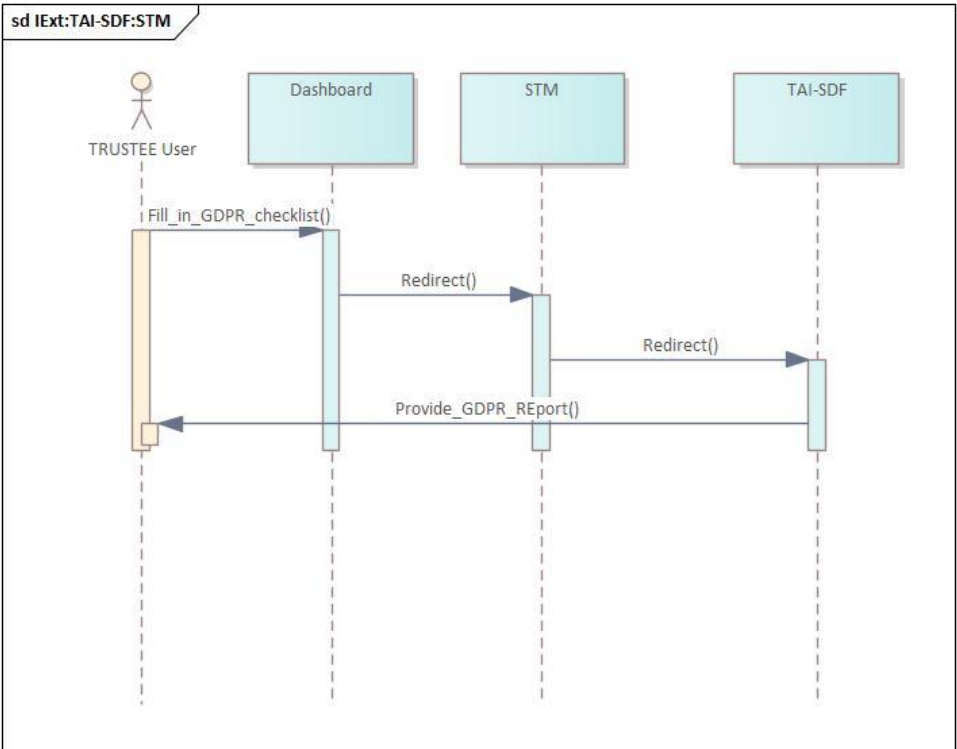


Figure 54: Sequence diagram related to interaction between GDPR compliance checker (inner module of TAI-SDF) and STM

5.5.3. TECHNICAL DESCRIPTION

5.5.3.1. UI DESCRIPTION

TAI-SDF has its own interface that has been developed according to the UI/UX guidelines commonly adopted at project level.

For the time being only GDPR compliance checker screens have been completed, in particular the screen related to the TAI-SDF survey module. The use case scenario related to the GDPR compliance checker is described in this document at section 5.5.2. When the Data Provider or Data Consumer- already authenticated and logged on the Dashboard- needs to fulfill the GDPR assessment, he/she will see the following screen:

Figure 55: screen related to GDPR compliance checker filling in questionnaire

Other screens aligned with UI/UX guidelines will be reported for future versions of this deliverable, namely D3.2 and D3.3.

5.5.3.2. DATA MODELS AND API DETAILS

As mentioned in section 5.5.2, TAI-SDF interacts with other ABB components of Trustee platform through the defined external interfaces, are the following:

- **Authentication Manager.** The TAI-SDF user (Data Entry) will need to verify their authentication with the Authentication Manager, as only authenticated users should be able to use the TRUSTEE tools.
- **AIMaaS.** The TRUSTEE user will be able to launch the TAI-SDF tool directly from the AIMaaS;
- **Accountable Transactions Recorder.** Activities, such as adding answers to questions, modifying answers to questions, and deleting answers to questions, will be recorded in ATR by sending a json with the following format:

Attribute	Type
userID	INTEGER
timestamps	INTEGER
actionType	INTEGER

{userID, timestamps, actionType}

- **Security and Trust Manager .** The TRUSTEE user will be redirected to the GDPR compliance checker module of TAI-SDF to perform self-assessment. Two jsons will be sent to STM, 1 json file with questions and 1 json file with answers. For each question the corresponding id, text, and the order in which the question is presented to the user are reported; the field "values" appears only in non- boolean, non free text questions as it follows:

```
[
  {
    "id": "1",
    "questionText": "Is the data you process (encrypted) personal data?"
  },
  {
    "id": "60",
    "values": [{"id": "1", "text": "Provider"}, {"id": "2", "text": "Consumer"}],
    "questionText": "Are you a Data Provider or a Data Consumer?"
  },
  ...
]
```

In the json file with answers, there is a reminder of the ids of the request and the compilation date, followed by the list of answers. For each answer there is a reference to the “id” in the json with questions; the field “value” means true/“false” for boolean questions, and the list of values for other questions (non-boolean and non-free text questions). Finally, the field “Status” reports “success/failed” for the status of the GDPR compliance assessment (“success” means that the report is completed, while “failed” means that the report is partially filled out or unavailable).

```
{
  "idDataset": "X",
  "idOwner": "X",
  "dateCreation": "2023-09-25",
  "Status": "Success"
  "answers":
  [
    {"idQuestion": "1", "value": "true"},
    {"idQuestion": "60", "value": "1,2"},
    ...
  ]
}
```

- **DPIA.** The PIA report will be shown to the TAI-SDF user in the privacy module.

At this stage only APIs related to GDPR compliance checker have been developed and implemented, that is APIs with authentication manager and STM.

5.6. VALIDATION

During the first implementation phase, Dry Run scenarios have been executed at RINA premises in order to test the core functionalities of the TAI-SDF from start to end. These scenarios aimed to validate the following functionalities:

- Project management chain: an AI project is created to start the trustworthiness assessment
- Checklist compiling process: Model/Data Providers or Consumers fill in checklists in order to assess the trustworthiness of the models/datasets they are providing or using
- Report generation: a report including the trustworthiness score and recommendations on how to improve such score is generated and presented to the Model/Data Providers or Consumers

Further functional tests will be reported in next versions of the deliverable, namely D3.2 and D3.3.

5.7. NEXT STEPS

The first version of TAI-SDF prototype includes the core functionalities, such as project and phase management, questionnaire compilation and report generation and only the finalised questionnaire for GDPR compliance checker.

The next version of the TAI-SDF will include:

- Finalised questionnaires of the inner TAI-SDF modules, such as Overall Framework, Fairness, Robustness, Explainability, Security and Privacy
- Improved user interface aligned with the latest version of UI/UX guidelines
- Implemented automatised mechanism to fill in questionnaires

The next phase of the TAI-SDF development, validation and integration will focus on the following areas:

- Test and validate functionalities
- Test the APIs with the other ABB components of TRUSTEE platform

5.8. CONCLUSIONS

This section provided an overview of the TAI-SDF component and inner modules. This ABB consists of different inner modules, a more general one Overall Framework that relates more to the general aspects that a user (Model/Data Provider and Consumer) has to consider during the research, analysis and study phase of its AI system and the other ones, each of them targeting the Trustworthiness principles, such as compliance, fairness, robustness, explainability, security and privacy. During the first implementation phase the focus was on the design of the questionnaires and on the definition of the architecture of TAI-SDF: a first standalone prototype has been implemented focusing on GDPR compliance. The interfaces with other TRUSTEE components have been investigated with definition of the corresponding data models. However, it is still a work in progress and further details will be provided in upcoming versions of this live document, namely D3.2 and D3.3.

6. PART IV: ACCOUNTABLE TRANSACTION RECORDER (ATR)

6.1. INTRODUCTION

The main aim of this task is to accomplish an Accountable Transaction Recorder (ATR) for TRUSTEE leveraging distributed ledger technology. The primary objective of this deliverable consists of defining functional, structural, and technical components, specifying user requirements, determining system specifications, outlining the overall architecture and marking out the legal requirements. These specifications will guide the research and development endeavour, tool development, demonstrations, and overall evaluation of the solution throughout the project lifecycle. The findings will help to establish the complete ATR architecture along with the interfaces, and integration requirements and also addressing the infrastructure enhancement, feature extraction configuration, and the design of explicit and implicit evaluation methods. This deliverable, within the scope of Task 3.3, provides an extensive analysis of distributed content management and data sharing models, examining state-of-the-art approaches and best practices. This work thus delivers guidelines to ensure the security, safety, and transparency of transaction recording methods within TRUSTEE. Additionally, the work carried out will help to determine the strengths and weaknesses of current approaches, identifies possible gaps, and highlights areas for improvement.

6.1.1. NEED AND OBJECTIVES

Interoperability refers to the capability that enables different systems, devices, applications to seamlessly communicate with each other. Interoperability often plays a crucial role by enabling consortiums to achieve greater efficiency. The ATR aims to communicate with other ABBs enabling seamless integration to record immutable and private transactions. The necessity of an ATR in the TRUSTEE project will serve the following reasons;

Transparency: The immutability nature of ATR will ensure tamper-resistant transactions; thus, ensuring full transparency in the recorded transactions.

Traceability: ATR will offer traceability of the recorded transactions; thus, enabling to achieve a transparent audit trail to track all the transaction history.

Security: The ATR will comprise cryptographic techniques where transactions are linked to each other forming blocks. The blocks of transactions are then linked to the adjacent blocks. This innovative process protects from unauthorised access and also ensures authenticity of the recorded transactions.

6.1.2. INNOVATION

The innovative features of the adopted blockchain will enhance the functionalities of ATR and enable it to provide improved features to the associated ABB's. The modularity will be one of the key design principles of ATR offering several unique approaches, such as customisable consensus, pluggable membership services, network partitioning, channels for privacy, etc. In addition to that, the private data collection provides an extra layer of security enabling the storage of confidential data with the functionalities of data deletion; thus, compliance with the GDPR. Consequently, leveraging the permissioned method, the ATR will comprise cutting-edge features which are nonexistent in other blockchain platforms.

6.2. BACKGROUND

6.2.1. BLOCKCHAIN TECHNOLOGY

Blockchain is a decentralised distributed ledger technology that can be leveraged to record transactions. It is an advanced storage mechanism that stores data in blocks which are then linked to each other forming a chain of blocks. Blockchain uses a cryptographic hash method which ensures data integrity, security and immutability (Sayeed 2019). While traditional database technologies comprise several challenges and are prone to various cyber-attacks, the immutable nature of blockchain and robust security attributes in place is a solution to various challenges. The blockchain can be public, private and permissioned. In a public blockchain, anyone can join to participate, whereas the private blockchain may give access to selected actors to be part of the network. Permissioned blockchain, on the other hand, is a private blockchain with strict access control and governance mechanisms in place. One core advantage of permissioned blockchain over the public is that it ensures rigid privacy and security and thus many security challenges such as 51% attack, n confirmation attack, and double-spending are less likely to disrupt the continuous flow of the network. Hyperledger Fabric, Hyperledger Besu, Corda, Ethereum Enterprise (EEA), and Celo are some of the permissioned blockchain platforms focusing on specific use cases and areas.

6.2.2. HYPERLEDGER FABRIC

Hyperledger Fabric is an open-source permissioned blockchain framework which focuses on providing a robust, secure and scalable distributed platform. Unlike public blockchains, Hyperledger emphasises privacy-preserving and allows to construct of a modular architecture thus enabling entities to construct the architecture according to their needs and requirements. The Hyperledger network often comprises several organisations or network members who can interact with each other to keep the regular flow of the network. Each organisation can have at least one or several peer nodes which come with distinct privileges (Nasir 2018).

Some of the major benefits of Hyperledger Fabric are as follows;

- Like many other blockchain solutions, Hyperledger does not require Byzantine fault Tolerance. Therefore, the efficient consensus mechanism enables it to offer enhanced performance when validating transactions.
- The formation of a Channel allows the network members to transact confidentially. Furthermore, to limit the exposure of confidential and sensitive data, it provides an additional layer of access control within the channel where access is limited to a subset of participants.
- The permissioned feature benefits every domain, such as healthcare, education, space, etc., to record transactions with full privacy and not be exposed to unknown parties.

6.2.3. CONSENSUS MECHANISM

A consensus is a method which ensures that all the blockchain operations are performed based on a standardised way. Consensus plays a key role in any blockchain network by enabling the network participants verify transactions and control overall security. Since blockchain does not comprise a central authority a consensus mechanism should be in place to maintain the regular flow of the network. Hyperledger Fabric provides a pluggable consensus approach thus the ordering of transactions is assigned to modular components that are separated from the peers. The modular aspect of consensus the blockchain platform to be Crash Fault-tolerance (CFT) or Byzantine Fault-Tolerance (BFT). Consensus defines the policy on how the peers would agree on the order and validity of the

transactions. There are various consensus mechanisms used by different blockchains which include Proof of Work (PoW), Proof of Stake (PoS), Delegated Proof of Stake (DPoS), Raft, Kafka, etc (Salimitari 2019).

6.2.4. CONTAINERISATION

The Hyperledger Fabric utilises container technology that enhances it to offer flexible services. The components of Hyperledger Fabric, such as the peer nodes, ordering service nodes (OSN), and chaincode are packaged and deployed as containers. The use of a container, such as Docker, enhances security, and scalability ensuring better resource management.

Docker is one of the most popular containerisation platforms that enables developers to generate, deploy and manage the containers thus maintaining the blockchain network and associated applications efficiently (Avram 2023).

The key advantages of leveraging the containerisation are:

- **Isolation:** The Docker containers provide a degree of isolation for the Hyperledger components ensuring there is no interference among the components and does not impact the host system. The degree of isolation confirms additional security and stability.
- **Portability:** The containers enfold the applications as well as the related dependencies thus making them portable in different environments. The applications can be executed on any compatible system without any modification.
- **Scalability:** Due to the flexibility in design, the containers can be scaled easily allowing easy replication and deployment of multiple instances to meet the changing demand.
- **Continuous Integration:** The containers are well-suited for contemporary software development methodologies ensuring a coherent building, testing and deployment of applications throughout various phases.

6.2.5. PUBLIC KEY INFRASTRUCTURE

The Public Key Infrastructure (PKI) comprises a collection of internet technologies that establish secure communications within an unsecure public network. PKI plays a crucial role enabling it to identify the network entities, devices, services and so on. It also provides controlled access to resources and thus ensures full security of data and accountability in the transactions. Some of the main components of PKI are x.509 digital certificates, public and private keys, certificate authorities, etc. In Hyperledger Fabric, the PKI ensures to offer a secure and trusted environment to manage identities and provide secure communications within the network (Yakubov 2018).

6.2.6. STATE DATABASE

The Hyperledger Fabric is designed to be compatible with both CouchDB and LevelDB; however, it does not support other database options such as PostgreSQL, MongoDB or MySQL. Although many approaches can be adopted to set up other databases, such as the state database, research indicates that to achieve pluggable database functionalities, a significant configuration would be required not only for the blockchain network but also for the database itself (Sharma 2018).

The **LevelDB** is a fast database which enables recording key-value pairs. It is a default database of fabric that can record the transaction logs of any application integrated to the ledger. LevelDB offers support solely for Key, Key range and composite key queries; for instance, Put(key-value), Get(Key, Value), and Delete (Key). Although LevelDB does not offer SQL queries, indexes or relational data models, some of the main advantages are: (i) It is a fast database with reduced overhead and less complex to utilise. (ii) It offers forward and backward interactions over the data.

CouchDB, on the other hand, is an open-source document-oriented database that stores information in JSON format. Besides LevelDB, the only alternative option Fabric offers is CouchDB. CouchDB is a NoSQL database enabling rich queries stored in JSON. Some of the main advantages of CouchDB are; (i) Making it convenient to meet the auditing and reporting requirements of Hyperledger Fabric by providing JSON queries and indices. (ii) Enabling the data to be exposed via HTTP URI thus offering to execute HTTP operations such as, GET, DELETE, PUT, POST against the data. (iii) making it convenient to query a large dataset of chaincode.

6.2.7. POLICIES

Policies in Hyperledger Fabric are a set of rules that defines how the decisions will be made in order to reach an outcome. Policies are one of the aspects which works differently in Hyperledger compared to other blockchain platforms, such as Ethereum or Bitcoin. Generally, in other blockchains any nodes can take part in the transaction generation and validation phase whereas Hyperledger being permissioned the entities need to be recognised by the underlying infrastructure and thus those entities can take part in deciding on the governance of the blockchain network before it is established and also make necessary changes after the network has been established (Soelman 2019).

The policies enable the channel organisations to reach any decisions on which network entity will have the authority to access or update the Fabric network and give the means to enforce such decisions. There are a various kind of policies which can be implemented in the Fabric network, some of the key policies as follows;

- **Endorsement Policies:** The chaincode package of the smart contracts defines the endorsement policy specifying the number of peers or organisations approvals are required to update a resource. Policies evaluate all the signatures associated with the transactions as well as proposals and verify whether the signatures meet the governance agreed by the network.
- **Modification Policies:** The Modification Policy is significant in the Hyperledger Fabric network because it defines how policies function and are updated in the network. It specifies which particular group of identities needed to approve the configuration update. Hence, the channel configuration comprises a reference to the policy that rules its modification.
- **Signature Policies:** The Signature policies identifies the sort of network entities required to sign for a policy to be satisfied. Signature policies are regarded as versatile due to the formation of extremely specific rules such as “An admin of Org B and 3 other admin, or 5 of 6 organisation admins”.

6.2.8. IMMUTABILITY

The immutability is the fundamental attribute of Hyperledger Fabric which will ensure the integrity and trustworthiness of the recorded transactions in the ATR. The moment a transaction gets validated and included in the chain, it becomes immutable and thus it cannot be deleted or altered. The ATR will achieve immutability utilising cryptographic hashing enabling each block to utilise a unique hash value to represent the recorded transactions. Any changes made to the transactions will produce a different hash value and thus evidence of the block being corrupted. The immutability feature of Fabric ensures a transparent and auditable history of transactions thus generating a tamper-resistant record of events (Wang 2018).

In contrast, the private or sensitive data from the “Private Data Collection” does not get stored in the shared ledger but rather maintained off-chain allowing access only to particular network participants. The immutability of the shared ledger is executed to the recorded transaction and the hash of the private data. Thus, the immutability of the private data is implemented through the adopted consensus mechanism and cryptographic hashing. As a result, the security and integrity of the private data stored in ATR can be fully ensured although the stored transactions are not available in the shared ledger.

6.2.9. PRIVATE DATA COLLECTION

ATR will comprise this functionality that will allow to keep confidential or private data separated thus giving access only to the authorised organisation to the transactional data. Since, creating a separate channel cause extra overhead, the private data collection feature will provide the blockchain organisations the ability to store data without creating a separate channel. Listing 1 defines the parameters and setting that require configuration to determine how private and sensitive data can be shared among the blockchain network participants (Benhamouda 2018).

The private data focuses on the following attributes:

- **requiredPeerCount:** defines the total number of peers needed to disseminate the private data as per the set policies.
- **maxPeerCount:** defines the total number of other peers, the particular endorsing peers responsible for circulating the data.
- **blockToLive:** defines the duration the private data will remain valid within the private database.
- **memberOnlyread:** defines whether the associated clients of the peer nodes are authorised to access the private data.

Listing 1Parameters show configuration details of private data collection.

```
[
  {
    "name": "collectionMarbles",
    "policy": "OR('Org1MSP.member', 'Org2MSP.member')",
    "requiredPeerCount": 0,
    "maxPeerCount": 3,
    "blockToLive":1000000,
    "memberOnlyRead": true,
    "memberOnlyWrite": true
  },
  {
    "name": "collectionMarblePrivateDetails",
    "policy": "OR('Org1MSP.member')",
    "requiredPeerCount": 0,
    "maxPeerCount": 3,
    "blockToLive":3,
    "memberOnlyRead": true,
    "memberOnlyWrite":true,
    "endorsementPolicy": {
      "signaturePolicy": "OR('Org1MSP.member')",
    }
  }
]
```

6.3. REQUIREMENTS

In this chapter, the compliance of the ATR ABB of TRUSTEE is investigated with regards to the user-centred, technical, legal, and socio-ethical requirements that were reported in D2.1. The results of this investigation provide further preliminary insight regarding with the aforementioned requirements addressed during the development and by the implementation and are presented in the following sub-sections. It is worth noting that only the Mandatory, Desirable, and Optional requirements have been considered at this point. Furthermore, only the currently compliant or partially-compliant requirements have been considered. As the implementation of the ATR progresses during the development of the TRUSTEE Platform and the maturity of the ABB increases, the investigation of the compliance will evolve and further updates and/or enhancements will be provided in upcoming versions of this deliverable, namely D3.2 and D3.3.

User-centred requirements were identified through the UX Research that was conducted under T2.3. These requirements were reported in D2.1 (TRUSTEE consortium, 2023), while they were also utilized in order to extract technical, functional and non-functional requirements, that were also outlined in D2.1. Table 14 below presents the interplay between user-centred requirements and technical requirements, that is the functional and non-functional requirements that address each user-centred requirement. The user-centred requirement ID is used to reference the respective requirement reported in D2.1 (TRUSTEE consortium, 2023) while the requirement's description is also

provided alongside whether the ATR is currently fully or partially compliant with the user-centred requirement, i.e., indicating whether technical requirements have been identified in order to address the respective user-centred requirement. The ID of the technical requirement, as reported in D2.1, that is addressing each user-centred requirement is also provided in the last column of the table. Additionally, it is possible than more than one technical requirement addresses a single user-centred requirement.

Table 14: ATR – User-centred requirements

Req id (WP2)	Req. Description	Compliance	How is addressed
USR-004	TRUSTEE must support more referential integrity in data modification.	Y	Req-ATR-FUNC-1, Req-ATR-FUNC-3, Req-ATR-FUNC-4, Req-ATR-FUNC-5
USR-005	The data provided must be used with respect to privacy regulations, legal frameworks, and ethical considerations and without the risk of data loss.	Y	Req-ATR-FUNC-1, Req-ATR-FUNC-2, Req-ATR-FUNC-3, Req-ATR-FUNC-5, Req-ATR-NFUNC-1
USR-039	TRUSTEE should be secure and free of cybersecurity issues.	P	Req-ATR-FUNC-1, Req-ATR-FUNC-3, Req-ATR-FUNC-4,
USR-057	TRUSTEE should be an efficient and highly secure platform	Y	Req-ATR-FUNC-2, Req-ATR-FUNC-3, Req-ATR-FUNC-5, Req-ATR-NFUNC-1

Table 15 presents the preliminary results of the investigation regarding the compliance of the ATR with the technical, functional and non-functional requirements specified in D2.1. The ID of the respective technical requirement is used as a reference to D2.1, while the requirement's description is also provided alongside whether DA complies with the requirement and how. Further updates and/or enhancements regarding the approach presented below for addressing each technical requirement will be provided in upcoming versions of this deliverable, namely D3.2 and D3.3.

Table 15: ATR – Functional and non-functional technical requirements

Req id (WP2)	Req. Description	Compliance	How is addressed
Req-ATR-FUNC-1	The decentralised network will consist of nodes each with distinct functionalities. The peers play a crucial role in validating, endorsing, and committing transactions. The nodes also take part in ledger management.	Y	Nodes will have distinct roles while peers will play a crucial role with transaction processing and ledger management. All trustee ABBs need to relate to peer nodes to record transactions.

Req-ATR-FUNC-2	Ordering service creates a consistent order of transactions before transactions are placed in the blockchain. The ordering service also ensures that all nodes comprise the same version of the ledger.	Y	The orderer is a crucial component of Hyperledger for transaction ordering and ledger synchronisation and will be implemented to meet the requirements.
Req-ATR-FUNC-3	A Membership Service Provider (MSP) ensures that only authorised participants can access and participate in the network. MSP defines policies for managing identity registration, authentication, authorization and so on.	Y	The implementation of MSP will ensure Identity registration, authentication, authorisation, and identity revocation.
Req-ATR-FUNC-4	The chaincode is responsible for executing all the smart contracts in the network. It contains the business logic and also defines how the infrastructure is going to behave, and what types of transactions will take place. Chaincode defines the rules and conditions that must be met in order to transactions to be validated. It can be designed to ensure privacy, for instance, private data collection, and it is executed in a secured environment to prevent malicious actors from tampering with the code.	Y	The chaincode will be installed in all the peer nodes. The chaincode goes through a lifecycle including installation, instantiation, and upgrade phases.
Req-ATR-FUNC-5	Hyperledger Explorer provides an interface to explore and interact with the blockchain. It is a web interface which provides graphical representation of the blockchain network. It can be utilised to determine transaction details, monitor performance, track network activity, etc.	Y	The monitoring tool will be configured to be integrated with the blockchain to obtain insights from ATR.
Req-ATR-FUNC-6	The main objective is to monitor the performance to detect anomalies. The integration with Hyperledger caliper will enable measuring all successful and failed transactions, transaction and read latency, transaction and read throughput, resource consumption, etc.	Y	Hyperledger caliper can be configured to integrate with Hyperledger in order to obtain performance metrics of ATR.

A preliminary analysis of the compliance with the legal and socio-ethical requirements defined in the D2.1 (TRUSTEE consortium, 2023) is shown in ANNEX 2 and ANNEX 3.

6.4. ARCHITECTURE

This section showcases both the high-level and low-level architecture of ATR. Figure 56 defines that the ATR is implemented using a single-channel blockchain network enabling an authorised TRUSTEE entity to record transactions and also make a query to the ledger. The ATR is composed of several network components, such as the peer nodes, orderers, MSP, CA and so on. Each component is responsible for performing particular tasks to maintain security and ensure an efficient recording of transactions. The peer nodes also comprise “private data collection” which is an off-chain storage enabling authorised entities to record sensitive and confidential data (if needed). The recorded information to the “private data collection” can be deleted if needed. The integration with Hyperledger Explorer enables ATR to obtain valuable network information from the blockchain.

Furthermore, a blockchain monitoring system is to be developed by TECNALIA and is properly detailed in D4.1. This component intends to provide to the project with monitoring capabilities over the blockchain of the project and therefore we considered that is important to mention it in the present deliverable. This monitoring system is a set of tools designed to observe, collect, and analyze data or metrics from the blockchain and provides real-time visibility into its performance, health, and behavior.

6.4.1. HIGH LEVEL

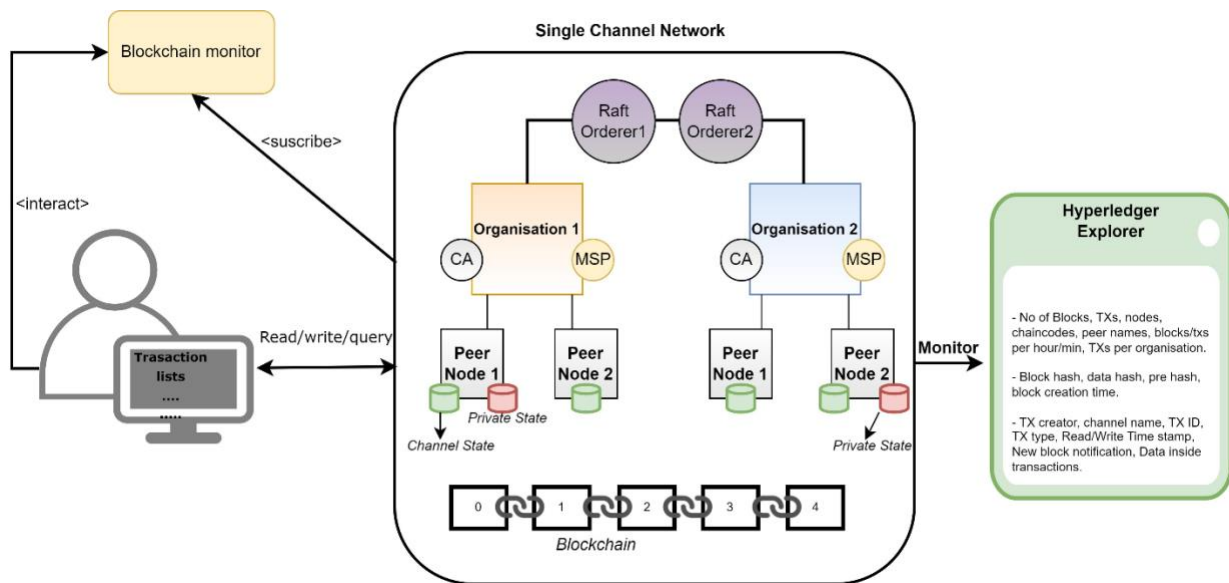


Figure 56 High level architecture of ATR established on permissioned blockchain with the monitoring system.

6.4.2. LOW LEVEL

Figure 57 shows the transaction recording phases at ATR implemented utilising Hyperledger Fabric. The recording of transactions to ATR goes through several phases before a transaction is considered confirmed. An authorised

TRUSTEE entity submits a transaction proposal to a peer node and a response is sent. The endorsing peers of the network take part in the endorsement process based on the defined policies and send the endorsement responses back. The endorsement responses are signed and sent to the orderers who are responsible for putting transactions in order, combining them into blocks and then sending them to all the peer nodes for committing. All committing peers verify the RW sets of transactions with the current world state and once committed, the transactions are recorded to the immutable ledger as appended in a new blockchain block.

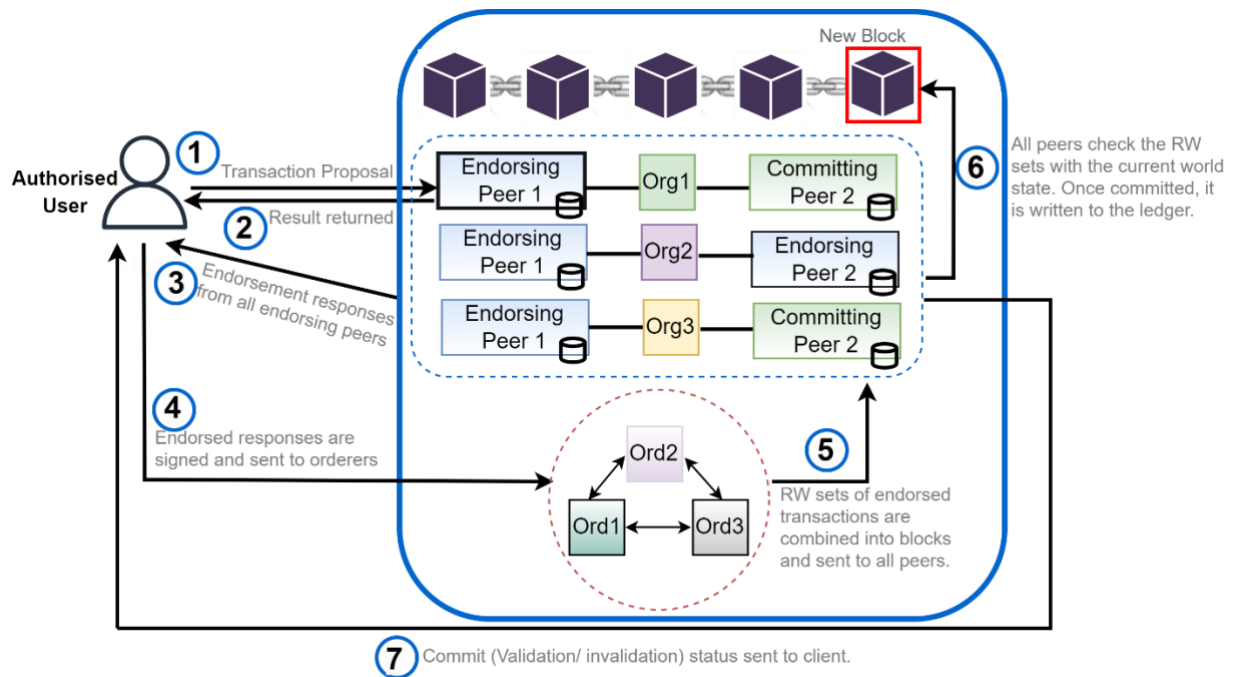


Figure 57 The transaction recording phases at ATR.

6.5. IMPLEMENTATION

During the first implementation phase, a permissioned blockchain network has been set up. All the network components have been established and the initial development of the chaincode enables to record of transactions to the ATR. In the next implementation phase, the implementation of the “private data collection” will be considered to permit the recording of sensitive data to ATR. The core functionalities that are operational are as follows;

6.5.1. FUNCTIONAL DESCRIPTION

The ATR will be implemented utilising Hyperledger Fabric that provides robust framework for distributed ledger technologies. The core functionalities of Hyperledger ensures a secure and scalable blockchain solution for recording transactions tailored to diverse needs of TRUSTEE platform. This section thus defines the core functionalities of the ATR that have been implemented to achieve the requirements;

1. **Channel:** The Hyperledger Fabric network can be constructed with one or several channels. The channel functions as a private “subnet” of communication among the network members where one or more peers can be part of it. Being permissioned, every entity intending to join must be authenticated and authorised

by the channel to perform the required task. While it is possible to comprise multiple channels where a peer node can be part of several channels and maintain several transaction recorders; however, multiple channels can cause unnecessary overheads, complications and channeling to manage; therefore, the ATR will be implemented within a single channel architecture providing efficient communication among the network members permitting them to execute both private and confidential transactions. The channel will be formed with various entities such as organisations, endorsing peers, committing peers, shared ledger, orderers and chaincode applications. Any transaction generated from TRUSTEE will be executed within the channel requiring the associated entities to serve with particular responsibilities to authenticate and authorise transactions.

2. **Chaincode:** Chaincode defines the logic based on the requirements set by the network participants. Chaincode is a software program which can be written in Go, Node.js, Java, etc. The chaincode can control the state of the ledger through the transactions submitted by other TRUSTEE ABB. The chaincode fully functions by managing the required logic that has been agreed upon by the entities involved. A chaincode can carry out various operations, such as it can invoke to write transactions, querying the ledger to obtain the recorded transactions and invoke another chaincode within the same channel or different channels.
Since chaincode executes logic agreed by all members of the blockchain network, they are required to sign it in order to prevent tampering. After a chaincode is endorsed, it can go through the installation and instantiation process. The network peers utilised Docker to run containers with chaincode in it.
3. **Ledger:** The ledger will be the main component responsible for storing and managing the immutable record of transactions. All the ATR transactions will be recorded to the ledger and thus serve the recorded information as tamper-resistant while ensuring full integrity and transparency. The ATR will utilise two types of ledgers offered by the Hyperledger Fabric;
 - World State Ledger: The world state consists of the current values; therefore, making it convenient for a program to directly retrieve the current value of the state rather than going through the whole transaction log. Since the ledger states are Key-value pairs, it offers flexibility while requiring to create, update and delete the states.
 - Transaction History Ledger or Blockchain: The transaction history ledger or the blockchain records every single change that occurs in the world state ledger. The transactions are accumulated within the blocks and appended to the blockchain thus enabling to determine the
4. **Peers:** Peer nodes are the primary element of the of the ledger for managing the ledgers and chaincodes. Since distinct version of Hyperledger Fabric can provide diverse functionalities, the Hyperledger Fabric v2.4 enables peer nodes to oversee transaction proposals as well as endorsements by operating the Fabric gateway service. Peers are the network nodes which are responsible for maintaining the blockchain ledger and executing transactions. Any entity that intends to write transactions to the ledger must be connected to a peer node of an organisation. The peers can be either endorsing peers or committing peers. The

endorsing peers are privileged peers with the ability to endorse transactions, while the committing peers are responsible for committing the endorsed and ordered transactions.

5. **Ordering Service Nodes (OSN):** The channel must have at least one or more Ordering Services Nodes (OSN) whose main task is to order the endorsed transactions into blocks. Unlike many public blockchains, the block size in Hyperledger Fabric can be configured based on the requirement. Thus, the number of transactions in a block can vary and the ordering service nodes or the orderers acquire the transactions from several applications. All the associated orders then arrange the batch of submitted transactions into sequence and then pancake them into blocks before sending to the committing peers.
6. **Membership Service Provider (MSP):** The Membership Service Provider (MSP) is a crucial component of Hyperledger Fabric. Its primary responsibility lies in managing the network identities, authentication and access control. The MSPs ensure that security and integrity are ensured through the provided services. The permissioned mechanism of Hyperledger Fabric requires every TRUSTEE entity to prove their identity to the channel before writing transactions to ATR. When an endorsing peer utilises their public key to sign a transaction, the MSP then takes charge of verifying whether the peer was authorised for transaction endorsement. Subsequently, the public key from the endorsing peer's certificate is further utilised to verify the validity of the signature appended to the transaction. Therefore, MSP is an important component of the permissioned network which ensures whether an identity can be trusted and recognised within the blockchain network.

Although named MSP, it is not responsible for providing anything to the blockchain network, instead, the requirement of the MSP is achieved by including a set of folders in the network configuration. The folders define MSP rules for the associated organisations within the blockchain network and other external organisations to validate the entities which are authorised to execute particular tasks.

To execute transactions to the ATR, a network member will need to perform the following tasks;

- Ensuring that the identity is created by a CA which is authorised by an organisation of the channel. The organisation's MSP establishes which CAs can be trusted by that organisation.
 - Ensuring that every organisation has their MSP appended to the channel thus making sure that the organisations are recognised and also approved by all the network entities.
 - Ensuring that the MSP is incorporated in the policy definition of the TRUSTEE blockchain network.
7. **Certificate Authority:** The Certificate Authorities (CA) is a trusted party responsible for creating public and private keys to generate key pairs to prove identity. Besides issuing certificates, it also takes charge of revoking certificates. Since there should also be a mechanism in place for the identities to be granted to the channel this is where the MSP takes charges to accomplish that. While deploying the ATR to the production environment, there could be two approaches TRUSTEE can adapt to serve the purposes;
 - While an organisation comprises its own CA, they can leverage that to issue the required certificates.
 - If no CA is available, they may utilise Hyperledger Fabric CA.
 8. **Purging Private Data:** Due to government regulations, it may be required to purge the sensitive data from the peer nodes. The unique feature of Hyperledger Fabric enables it to purge sensitive or private data so

that it is not obtainable through chaincode query or to peers that are not authorised to request the data. The private data can be removed from the state the same way as regular state data thus ensuring that the stored data cannot be accessed. It should be noted that by simply deleting sensitive data from the private, the history of the private data still endures in the node's private database. To remove the private/sensitive data entirely the chaincode API `PurgePrivateData` must be executed instead of `DelPrivateData` API.

Functionality	Description	Functional
Blockchain network	A Single channel ledger.	Yes
Network Components	The network components, such as the peer nodes, ordering service nodes, organisations, has been established.	Yes
Chaincode	The initial development of the chaincode enables to record and query transactions.	Yes
Blockchain Monitoring	Hyperledger Explorer is a monitoring tool enabling to monitor internal ATR services	Yes
Private Data Collection	Enabling TRUSTEE ABBs to record sensitive data	Partially

Figure 58 shows the sequence diagram allowing any authorised TRUSTEE ABB to interact to record transactions to ATR. While the response is sent back to the ABBs, the recorded transactions can be viewed from the Dashboard.

6.5.2. OPERATIONAL DESCRIPTION

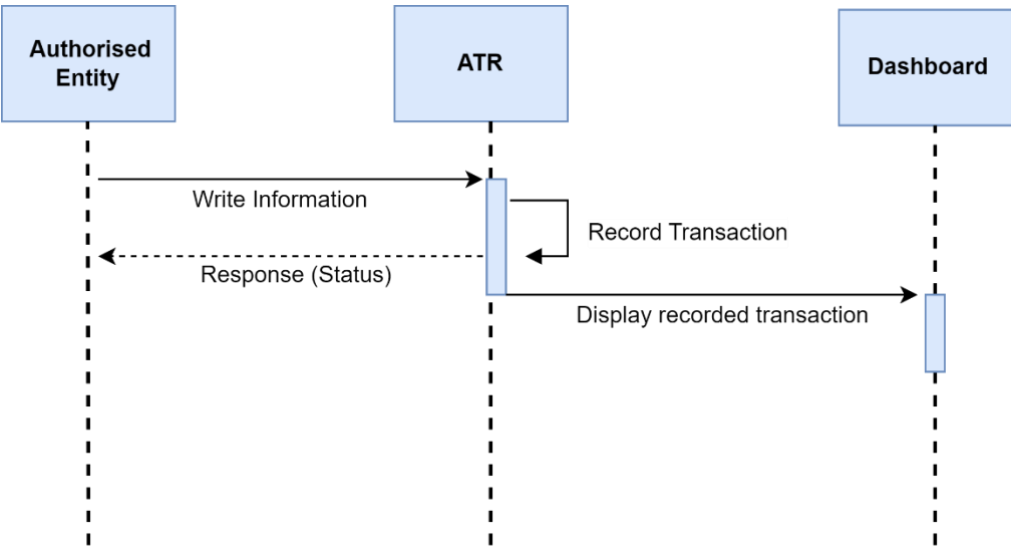


Figure 58 Sequence diagram showing the interactions to ATR.

6.5.3. TECHNICAL DESCRIPTION

In this section, a comprehensive overview of the project's implementation details, covering Input/Output/Data Models, Graphical Interface (including mock-ups and frontend components), and the API will be presented. It highlights the technology stack, programming languages, and development methodologies utilized in the project's construction, along with data flow, user interactions, and data validation mechanisms.

6.5.4. TECHNICAL DETAILS

The Technical Details section provides an in-depth overview of the development, installation, and deployment aspects of the TRUSTEE platform. This section is crucial for technical stakeholders and development teams to understand the tools, technologies, and infrastructure required to successfully implement and deploy the ATR ABB. Table 16 gathers the implementation requirements and Listing 2 shows a generic structure that is considered for communication from TRUSTEE ABBs and applied for the first phase of implementation to record transactions.

Table 16 Technical details for the first phase of implementation.

DEVELOPMENT	
Type (Software/Hardware)	Software
Operating Systems	Linux (Ubuntu)
Database and databases tools	LevelDB, CouchDB
GUI	No

Programming language	Node.js (JavaScript) or Go
Development Environment	Visual Studio Code
INSTALLATION AND DEPLOYMENT	
Software Requirements	Docker, Hyperledger Fabric
Hardware Requirements	Standard Hardware (e.g., 16GB RAM)
Containerization	Docker
Communications	gRPC

Listing 2 Communication between TRUSTEE ABB and ATR for recording transactions.

```
{
  "transactionID": "tx123",
  "providerID": "STM123",
  "description": "This is private data",
  "transaction_field1": "Value1",
  "transaction_field2": "Value2",
  "transaction_field3": "Value3",
  "transaction_field4": "Value4"
}
```

6.6. VALIDATION

6.6.1. CHANNEL CREATION AND NETWORK COMPONENTS

- Figure 59 shows the creation of a single channel called ‘trustee’ and the active network components, such as peer nodes, orderers, CA, etc., running on Docker container.

```
+ configtxlator proto_encode --input config_update_in_envelope.json --type common.Envelope --output Org2MSPanchors.tx
2023-10-22 00:26:50.048 UTC 0001 INFO [channelCmd] InitCmdFactory -> Endorser and orderer connections initialized
2023-10-22 00:26:50.705 UTC 0002 INFO [channelCmd] update -> Successfully submitted channel update
Anchor peer set for org 'Org2MSP' on channel 'trustee'
Channel 'trustee' joined
sarwar@experiment:~/Desktop/ATR/fabric-samples/test-network$ docker ps
```

CONTAINER ID	IMAGE	COMMAND	CREATED	NAMES	STATUS	PORTS
1110313db0ac	hyperledger/fabric-tools:latest	"/bin/bash"	39 seconds ago	cli	Up 36 seconds	
480cf3bdea94	hyperledger/fabric-peer:latest	"peer node start"	41 seconds ago	peer0.org2.example.com	Up 38 seconds	0.0.0.0:9051->9051/tcp, :::9051->9051/tcp
d6dbcbfadf74	hyperledger/fabric-peer:latest	"peer node start"	41 seconds ago	peer0.org1.example.com	Up 38 seconds	0.0.0.0:7051->7051/tcp, :::7051->7051/tcp
p, 0.0.0.0:9444->9444/tcp, :::9444->9444/tcp	hyperledger/fabric-orderer:latest	"orderer"	41 seconds ago	orderer.example.com	Up 38 seconds	0.0.0.0:7050->7050/tcp, :::7050->7050/tcp
p, 0.0.0.0:7053->7053/tcp, :::7053->7053/tcp, 0.0.0.0:9443->9443/tcp, :::9443->9443/tcp	hyperledger/fabric-ca:latest	"sh -c 'fabric-ca-se...'"	About a minute ago	ca_orderer	Up About a minute	0.0.0.0:9054->9054/tcp, :::9054->9054/tcp
p, 7054/tcp, 0.0.0.0:19054->19054/tcp, :::19054->19054/tcp	hyperledger/fabric-ca:latest	"sh -c 'fabric-ca-se...'"	About a minute ago	ca_orderer	Up About a minute	0.0.0.0:7054->7054/tcp, :::7054->7054/tcp
5c4f041861c9	hyperledger/fabric-ca:latest	"sh -c 'fabric-ca-se...'"	About a minute ago	ca_org1	Up About a minute	0.0.0.0:8054->8054/tcp, :::8054->8054/tcp
p, 0.0.0.0:17054->17054/tcp, :::17054->17054/tcp	hyperledger/fabric-ca:latest	"sh -c 'fabric-ca-se...'"	About a minute ago	ca_org2	Up About a minute	0.0.0.0:8054->8054/tcp, :::8054->8054/tcp
8510735e8a02	hyperledger/fabric-ca:latest	"sh -c 'fabric-ca-se...'"	About a minute ago	ca_org2	Up About a minute	0.0.0.0:8054->8054/tcp, :::8054->8054/tcp
p, 7054/tcp, 0.0.0.0:18054->18054/tcp, :::18054->18054/tcp						

Figure 59 creation of single channel and network services running on docker.

- Figure 60 shows the committing of chaincodes at the peer nodes and running on top of the container along with other network services.

```
2023-10-22 02:10:48.280 BST 0001 INFO [chaincodeCmd] ClientWait -> txid [bfa7b273360e1616e91ef48a2ce77a909fbb174757d0a1af8b88fe735ccd5d78] committed with status (VALID) at localhost:9051
2023-10-22 02:10:48.212 BST 0002 INFO [chaincodeCmd] ClientWait -> txid [bfa7b273360e1616e91ef48a2ce77a909fbb174757d0a1af8b88fe735ccd5d78] committed with status (VALID) at localhost:7051
sarwar@experiment:~/Desktop/ATR/fabric-samples/test-network$ docker ps
```

CONTAINER ID	IMAGE	COMMAND	CREATED	NAMES	STATUS	PORTS
9a30dd9cbab4	dev-peer0.org1.example.com-basic_1.0-7a5ded0f6679ada05df35b2ea20b3b59fd841564f4b831b57664dfc1f2543c3-bccfea2c8e1f591e08dbcb741fe6a04530ac48d75978ff801512a1f55e4a482c	"docker-entrypoint.s..."	2 minutes ago	dev-peer0.org1.example.com-basic_1.0-7a5ded0f6679ada05df35b2ea20b3b59fd841564f4b831b57664dfc1f2543c3	Up 2 minutes	
d2e48a3ffa81	dev-peer0.org2.example.com-basic_1.0-7a5ded0f6679ada05df35b2ea20b3b59fd841564f4b831b57664dfc1f2543c3-3d5c5022bb980e8c8faaa2a936051d6b2119c5805cbd0156076a9efb91712e86	"docker-entrypoint.s..."	2 minutes ago	dev-peer0.org2.example.com-basic_1.0-7a5ded0f6679ada05df35b2ea20b3b59fd841564f4b831b57664dfc1f2543c3	Up 2 minutes	
1110313db0ac	hyperledger/fabric-tools:latest	"/bin/bash"	46 minutes ago	cli	Up 46 minutes	
480cf3bdea94	hyperledger/fabric-peer:latest	"peer node start"	46 minutes ago	peer0.org2.example.com	Up 46 minutes	0.0.0.0:9051->9051/tcp, :::9051->9051/tcp, 7051/tcp, 0.0.0.0:9445->9445/tcp, :::9445->9445/tcp
d6dbcbfadf74	hyperledger/fabric-peer:latest	"peer node start"	46 minutes ago	peer0.org1.example.com	Up 46 minutes	0.0.0.0:7051->7051/tcp, :::7051->7051/tcp, 0.0.0.0:9444->9444/tcp, :::9444->9444/tcp
44->9444/tcp	hyperledger/fabric-orderer:latest	"orderer"	46 minutes ago	orderer.example.com	Up 46 minutes	0.0.0.0:7050->7050/tcp, :::7050->7050/tcp, 0.0.0.0:7053->7053/tcp, :::7053->7053/tcp
53->7053/tcp, 0.0.0.0:9443->9443/tcp, :::9443->9443/tcp	hyperledger/fabric-ca:latest	"sh -c 'fabric-ca-se...'"	46 minutes ago	ca_orderer	Up 46 minutes	0.0.0.0:9054->9054/tcp, :::9054->9054/tcp, 7054/tcp, 0.0.0.0:19054->19054/tcp, :::19054->19054/tcp
fdefb7d13262	hyperledger/fabric-ca:latest	"sh -c 'fabric-ca-se...'"	46 minutes ago	ca_orderer	Up 46 minutes	0.0.0.0:7054->7054/tcp, :::7054->7054/tcp, 0.0.0.0:17054->17054/tcp, :::17054->17054/tcp
4/tcp, :::19054->19054/tcp	hyperledger/fabric-ca:latest	"sh -c 'fabric-ca-se...'"	46 minutes ago	ca_org1	Up 46 minutes	0.0.0.0:8054->8054/tcp, :::8054->8054/tcp, 7054/tcp, 0.0.0.0:18054->18054/tcp, :::18054->18054/tcp
5c4f041861c9	hyperledger/fabric-ca:latest	"sh -c 'fabric-ca-se...'"	46 minutes ago	ca_org1	Up 46 minutes	0.0.0.0:8054->8054/tcp, :::8054->8054/tcp, 7054/tcp, 0.0.0.0:18054->18054/tcp, :::18054->18054/tcp
17054->17054/tcp	hyperledger/fabric-ca:latest	"sh -c 'fabric-ca-se...'"	46 minutes ago	ca_org2	Up 46 minutes	0.0.0.0:8054->8054/tcp, :::8054->8054/tcp, 7054/tcp, 0.0.0.0:18054->18054/tcp, :::18054->18054/tcp
8510735e8a02	hyperledger/fabric-ca:latest	"sh -c 'fabric-ca-se...'"	46 minutes ago	ca_org2	Up 46 minutes	0.0.0.0:8054->8054/tcp, :::8054->8054/tcp, 7054/tcp, 0.0.0.0:18054->18054/tcp, :::18054->18054/tcp
4/tcp, :::18054->18054/tcp						

Figure 60 Committing of chaincode on the peer nodes.

- Figure 61 shows the chaincode is invoked to record transactions and the query made to the ledger to list the recorded transactions. The provided timestamp is generated from the ledger during the endorsement process.

```
sarwar@experiment:~/Desktop/ATR/fabric-samples/test-network$ peer chaincode invoke -o localhost:7050 --ordererTLSHostnameOverride order
er.example.com \
--tls --cafile $ORDERER_CERTFILE \
-C trustee -n basic \
--peerAddresses localhost:7051 --tlsRootCertFiles $ORG1_PEER_CERTFILE \
--peerAddresses localhost:9051 --tlsRootCertFiles $ORG2_PEER_CERTFILE \
-c '{"function": "storeTransaction", "Args": ["tx123", "KR123", "Storing Content Identifier", "QmPK1s3pNYL19ERiq3BDxKa4XosgWwFRQUydHU
tz4YgpqB", "SMJq1L1GL19ERiq3BDxKa4XosgWwFRQUydHUTz4HgTR", "BhkR1s3pNYL19ERiq3BDxKa4XosgWwFRQUydHUTz4jRt18", "Y7Gp0s3pNYL19ERiq3BDxKa4Xo
sgWwFRQUydHUTz9uyIoG"]}'
2023-10-22 02:28:12.484 BST DB01 INFO [chaincodeCmd] chaincodeInvokeOrQuery -> Chaincode invoke successful. result: status:200 payload:
{"id":"tx123","timestamp":"1697938092","providerID":"KR123","description":"Storing Content Identifier","field1":"QmPK1s3pNYL19ERiq3BDxKa4XosgWwFRQUydHUTz4YgpqB","
field2":"SMJq1L1GL19ERiq3BDxKa4XosgWwFRQUydHUTz4HgTR","field3":"BhkR1s3pNYL19ERiq3BDxKa4XosgWwFRQUydHUTz4jRt18","field4":"Y7Gp0s3pNYL19ERiq3BDxKa4XosgWwFRQUydHUTz9uyIoG"}
sarwar@experiment:~/Desktop/ATR/fabric-samples/test-network$ peer chaincode query -C trustee -n basic -c '{"Args":["queryAllCIDsAndMeta
data"]}' | jq
[
  {
    "id": "tx123",
    "timestamp": "1697938092",
    "providerID": "KR123",
    "description": "Storing Content Identifier",
    "field1": "QmPK1s3pNYL19ERiq3BDxKa4XosgWwFRQUydHUTz4YgpqB",
    "field2": "SMJq1L1GL19ERiq3BDxKa4XosgWwFRQUydHUTz4HgTR",
    "field3": "BhkR1s3pNYL19ERiq3BDxKa4XosgWwFRQUydHUTz4jRt18",
    "field4": "Y7Gp0s3pNYL19ERiq3BDxKa4XosgWwFRQUydHUTz9uyIoG"
  }
]
```

Figure 61 Transaction recording and query to the ledger

- Figure 62 shows endorsement failure during invoke which resulted in an error response due to having the same transaction in the ledger with the txid.

```
sarwar@experiment:~/Desktop/ATR/fabric-samples/test-network$ peer chaincode query -C trustee -n basic -c '{"Args":["queryAllCIDsAndMeta
data"]}' | jq
[
  {
    "id": "tx123",
    "timestamp": "1697938092",
    "providerID": "KR123",
    "description": "Storing Content Identifier",
    "field1": "QmPK1s3pNYL19ERiq3BDxKa4XosgWwFRQUydHUTz4YgpqB",
    "field2": "SMJq1L1GL19ERiq3BDxKa4XosgWwFRQUydHUTz4HgTR",
    "field3": "BhkR1s3pNYL19ERiq3BDxKa4XosgWwFRQUydHUTz4jRt18",
    "field4": "Y7Gp0s3pNYL19ERiq3BDxKa4XosgWwFRQUydHUTz9uyIoG"
  }
]
sarwar@experiment:~/Desktop/ATR/fabric-samples/test-network$ peer chaincode invoke -o localhost:7050 --ordererTLSHostnameOverride order
er.example.com \
--tls --cafile $ORDERER_CERTFILE \
-C trustee -n basic \
--peerAddresses localhost:7051 --tlsRootCertFiles $ORG1_PEER_CERTFILE \
--peerAddresses localhost:9051 --tlsRootCertFiles $ORG2_PEER_CERTFILE \
-c '{"function": "storeTransaction", "Args": ["tx123", "KR123", "Storing Content Identifier", "QmPK1s3pNYL19ERiq3BDxKa4XosgWwFRQUydHU
tz4YgpqB", "SMJq1L1GL19ERiq3BDxKa4XosgWwFRQUydHUTz4HgTR", "BhkR1s3pNYL19ERiq3BDxKa4XosgWwFRQUydHUTz4jRt18", "Y7Gp0s3pNYL19ERiq3BDxKa4Xo
sgWwFRQUydHUTz9uyIoG"]}'
Error: endorsement failure during invoke. response: status:500 message:"Transaction: tx123 exists"
sarwar@experiment:~/Desktop/ATR/fabric-samples/test-network$
```

Figure 62 Endorsement of transaction failure due to duplicate transaction.

- Figure 63 shows a query made to the ledger listing all the recorded transactions. While the 'id' and the 'provider id' remain mandatory, the timestamp is the endorsement timestamp generated from the ledger. The Fields have been populated by the TRUSTEE ABBs according to their needs and thus some fields remain null.

```

sarwar@experiment: ~/Desktop/ATR/fabric-samples/test-network$ peer chaincode query -C trustee -n basic -c '{"Args":["queryAllCIDsAndMetadata"]}'
[
  {
    "id": "tx123",
    "timestamp": "1697938092",
    "providerID": "KR123",
    "description": "Storing Content Identifier",
    "field1": "QmPK1s3pNYL19ERlq3BDxKa4XosgHwFRQyudHutz4YgpgB",
    "field2": "5WJq111GL19ERlq3BDxKa4XosgHwFRQyudHutz4HgTR",
    "field3": "8hkR1s3pNYL19ERlq3BDxKa4XosgHwFRQyudHutz4jRt18",
    "field4": "Y7Gp0s3pNYL19ERlq3BDxKa4XosgHwFRQyudHutz9uyIoG"
  },
  {
    "id": "tx1234",
    "timestamp": "1697938692",
    "providerID": "KR1234",
    "description": "Storing Content Identifier",
    "field1": "JLM71s3pNYL19ERlq3BDxKa4XosgHwFRQyudHutz6trFG0",
    "field2": "KdAz8L1GL19ERlq3BDxKa4XosgHwFRQyudHujSwe60P",
    "field3": "",
    "field4": ""
  },
  {
    "id": "tx12345",
    "timestamp": "1697939034",
    "providerID": "STM123",
    "description": "Storing Agreement",
    "field1": "Agreement ID: ABC-123",
    "field2": "Dataset ID: CBA-123",
    "field3": "OrganisationID: ACD-123",
    "field4": "Agreement Status"
  },
  {
    "id": "tx123456",
    "timestamp": "1697939511",
    "providerID": "STM1234",
    "description": "Storing Agreement",
    "field1": "Agreement ID: ABC-567",
    "field2": "Dataset ID: CBA-456",
    "field3": "OrganisationID: ACD-4560",
    "field4": "Agreement Status"
  },
  {
    "id": "tx1234567",
    "timestamp": "1697939567",
    "providerID": "TAI-SDF123",
    "description": "",
    "field1": "Action Type ",
    "field2": "",
    "field3": "",
    "field4": ""
  }
]

```

Figure 63 Query to list all transactions recorded by TRUSTEE ABBs.

6.7. NEXT STEPS

The next phase of the ATR development will focus on the following areas;

- Implementation of a “Private Data Collection” feature allowing TRUSTEE ABBs to record sensitive data.
- Integrate with KR enabling KR to store Content Identifier to ATR.
- Carry out the performance analysis of ATR.

6.8. CONCLUSIONS

The main functionality of ATR involves recording transactions sent from TRUSTEE ABBs. ATR is implemented utilising private permissioned blockchain and the immutability feature of the adopted technology guarantees the reliability and trustworthiness of the recorded transactions. In addition, the permissioned nature of blockchain technology also ensures enhanced privacy and confidentiality. Unlike many blockchain solutions, ATR implementation will comprise a “Private Data Collection” feature, a unique method that allows entities to record sensitive data off-chain thus providing full flexibility when it is necessary to delete data. This document has showcased the major functionalities of ATR. It has demonstrated how ATR empowers permitted ABBs to record transactional data and enables them to search transactions based on the TX-ID and return errors when duplicate transactions are detected. For future implementation, the main focus will be implementing the “Private Data Collection” feature, also improving the task carried out to align with the requirements.

7. PART V: KNOWLEDGE REPOSITORY (KR)

7.1. INTRODUCTION

The main purpose of this chapter is to analyse the architecture of the Knowledge Repository (KR) and indicate the technologies required to develop, implement, and utilise a distributed peer-to-peer content-sharing protocol. The knowledge repository aims to store knowledge-related content stemming from search queries, results of computations over the encrypted data of Data Providers, and information about AI/ML use and/or inference run results. It will be implemented utilising a decentralised distributed method which is highly resilient, scalable and fault tolerant.

This document thus delivers guidelines to ensure secure storage of TRUSTEE knowledge contents in the distributed system. It will commence with background information providing an overview of the architecture and insights into the technologies adopted. Additionally, this work demonstrates the functionalities, identifies possible gaps, and highlights areas for improvement along with future implementation.

7.1.1. NEED AND OBJECTIVES

The TRUSTEE Knowledge Repository will follow a FAIR model and aggregate particulars about the structure of the content, associated information about its sources and social connectivity. It is therefore essential to construct a repository which is resilient and fault-tolerant for the secure distribution of knowledge-related content. Furthermore, the efficiency and fast retrieval of the knowledge-based content, consistent content availability and content redundancy also need to be ensured at all times. Hence, the knowledge repository will provide the integrity, trust, and security of any content stored within it.

To meet the requirements, several key factors are needed that will enhance the functionalities of KR, including;

Decentralisation: The traditional repositories are primarily based on centralised servers thus prone to various attacks and loss of content occurs due to a single point of failure. Therefore, it is necessary to leverage a distributed approach which will boost resilience and ensure fault-tolerance.

Content Redundancy and Availability: It is crucial to avoid storing redundant knowledge contents in the repository and thus a mechanism should be in place to identify contents based on their uniqueness. As a result, TRUSTEE KR must have a mechanism in place to ensure no identical contents are stored to boost content redundancy and availability.

Efficient Content Management: A single or centralised repository might struggle to meet the TRUSTEE requirements, so to improve the knowledge access times and reduce the load, a distributed hash table (DHT) method should be in place to enhance the efficiency in retrieving knowledge-related contents from distributed nodes.

The primary objectives of this deliverable are as follows;

- **Distributed Repository:** The primary objective is to design and implement a Knowledge Repository (KR) which will fully align with TRUSTEE reference architecture and entirely comply with the requirements. It should also provide Knowledge Graph and analysis functionalities to gain deep insight from the stored contents.

- **Interoperability:** The KR should be able to integrate with various TRUSTEE ABBs, such as the ATR. It should be flexible to promote interoperability and enhance continuous content sharing among different networks.
- **Content Integrity and Trust:** It is necessary to ensure the integrity of the stored knowledge contents and enhance trust. Thus, a cryptographic hashing mechanism will be in place to identify and verify contents.
- **Knowledge Management:** The KR aims to perform effective knowledge management and version control of the stored knowledge. To ensure complete privacy and security, the KR should ensure the immutability of the stored knowledge while at the same time providing the content deletion functionality.

7.1.2. INNOVATION

Trustee KR will be implemented utilising the InterPlanetary File System (IPFS). The innovative features of IPFS will enhance the functionalities of the KR and enable it to meet all the objectives. IPFS will ensure that TRUSTEE KR achieves redundancy of the stored knowledge with the ability to content deletion and offline access. Furthermore, secure communication among nodes and immutability of the knowledge-based contents will also be established to obtain content integrity and security.

7.2. BACKGROUND

7.2.1. INTERPLANETARY FILE SYSTEM (IPFS)

The InterPlanetary File System (IPFS) is a peer-to-peer protocol built by Protocol Labs. It is a distributed system which comprises similar functionalities as a centralised cloud, however based on decentralised architecture. IPFS provides decentralised service and therefore can comprise several distributed systems to host the knowledge-based content. The hosted contents can range from databases to websites, media files, documents, apps, AI models, search queries and so on. A user can access any knowledge-related content from the repository, hosted in the IPFS network, with a content identifier (CID). IPFS functions similarly to a blockchain network where several nodes offer their storage bandwidth. The public IPFS network is formed of hundreds of thousands of nodes being open to anyone, whereas the private IPFS provides access to specific groups requiring authorisation to join (Benet 2014).

7.2.2. PRIVATE IPFS NETWORK

The private IPFS network provides enhanced security compared to public IPFS allowing access only to the permissioned nodes. In a private IPFS, the access is restricted to only authorised users. Besides ensuring enhanced security, it allows secure collaboration among authorised entities without the risk of content being compromised. Furthermore, it can also permit custom naming services enabling to access knowledge-related content through human-readable names. Although the private IPFS offer several benefits, the public IPFS takes advantage of the massive scaling, providing geographic resiliency (Manoj 2020). In terms of functionalities, the private IPFS performs the same way as the public, but the nodes are only permitted to communicate with their peers within that particular network. To join an IPFS network, the nodes are provided with a secret key, which is also referred to as a “Swarm key”. The Swarm key comes in a text file consisting of the secret needed for peers intending to join the IPFS private instance enabling the peer nodes with the same swarm key to communicate with each other.

```

/ key / swarm / psk /1.0.0/

/ base16 /

8981939525395 dc0ed49ffc8c9d2e91b6e23bb591ccc53762e8a4e83946d5f5

```

Listing 3 An example of swarm key

7.2.3. PRIVATE IPFS ADVANTAGES

The private KR, implemented in an IPFS network, provides several advantages when compared to public repositories;

- **Enhanced privacy:** In a private network, the knowledge contents can be stored and shared only with authorised users thus confirming enhanced privacy and confidentiality.
- **Knowledge security:** The private KR can provide greater flexibility to impose various security measures including encryption, access controls, authentication, etc to secure the knowledge resource from unauthorised access.
- **Faster Knowledge Retrieval:** The private network can enable quick access to the stored knowledge due to lower latency and faster resource retrieval times.
- **Cost Management:** Compared to a public repository, the private KR can optimise the cost associated the knowledge storage and network operations.
- **Flexible Network Topologies:** The private KR network can be customised to align with requirements, and therefore the number of nodes, dedicated gateways and other network parameters can be customised.
- **Compliance and regulatory Alignment:** The flexible configuration approach can enable the KR to comply with industry-specific regulations and compliance standards. This ensures enhanced flexibility to entities that comprise strict regulatory requirements.

7.2.4. CONTENT-ADDRESSABLE STORAGE (CAS)

The content-addressable storage (CAS) utilises hashes as reference enabling one to obtain the benefits of quoting while maintaining constant reference size, regardless of the size of the content. IPFS will enable KR to comprise a content addressable storage approach to identify and acquire the stored knowledge contents. Unlike traditional systems, the CAS does not address the contents based on their location but rather based on the generated hash. A cryptographic hash, such as SHA-256, is used which acts as an identifier of the content. The CAS ensures efficient file retrieval as identical contents will always generate the same identifier while evidencing if content has been tampered with.

7.2.5. DISTRIBUTED HASH TABLE (DHT)

The Distributed Hash Table (DHT) is a key value look-up table which links the CID to the locations of the content. The DHT is leveraged for routing and informing peers about the location of the resource. Every peer in the

distributed KR network will maintain a list of the known peers and update it when new peers join the network. The DHT table utilised within KR by IPFS is based on Kademlia which is widely used in p2p systems. Generally, when accessing and sharing knowledge resources, the DHT can be queried directly instead of high-level add, get and cat tools. The below helps to locate peers that can offer a particular bit of the stored content.

```
ipfs dht findprovs QmW2WQi7j6c7UgJTArActp7tDNikE4B2qXtFCfLPdsgaTQ
```

Listing 4 An example of locating content from DHT

7.2.6. LIBP2P

Libp2p is an open-source modular networking stack and protocol suite developed by Protocol Labs. It is a crucial component of IPFS and other decentralised networks. The Libp2p provides a flexible, modular, and agnostic networking framework to enhance secure and efficient communications among the peer nodes. After a CID is determined, the next phase is the discovery phase which utilises the DHT for mapping keys to values. The DHT functions in a distributed manner with IPFS peers that are looking for help from the neighbours to determine the mapping. Libp2p manages the communications between hosts for peer interactions. Initially, the requestor discovers the peers that are hosting a knowledge resource and then the DHT is effective again to determine the current location of these peers. The process of both querying and locating the contents leverages the libp2p protocol.

7.2.7. MERKLEDAG (DIRECTED ACYCLIC GRAPH)

A Merkle DAG comprises similar functionalities to a MERKLE tree, but it does not require to be balanced which means it is not mandatory to obtain an equal number of child nodes. In contrast to the Merkle tree, Merkle DAG enables every node to carry a payload. In a Merkle DAG, each node comprises an identifier that is the outcome of the contents in the form of a hash.

Some key attributes that Merkle DAG comprises include;

- Merkle DAGS can only be constructed from the leaves of the nodes with no children. The parent nodes are appended after the children's because the identifiers of the children are required to be computed before linking them.
- All nodes within the Merkle DAG serve as the root of its own (sub) Merkle DAG and thus this subgraph is encompassed within the parent DAG.
- All Merkle DAG nodes comprise immutability. Therefore, changes made in a node deflect the identifier affecting the DAG by generating a completely new DAG.

7.2.8. CONTENT IDENTIFIER (CID)

A content identifier is a hash which can be used to reference a knowledge resource in the repository. Although a CID does not provide information on the location of the stored content, it creates some form of address based on the content itself. The CIDs are concise regardless of the size of the stored knowledge content. The KR will utilise the SHA-256 hashing algorithm and support other algorithms. Every content, such as knowledge-related resource, regardless of the format, shared within the distributed network can be identified by the CID. Since the CID is

acquired directly from the resource it will always be the same regardless of the file name or any other associated metadata. This ensures that the correct content is obtained from the network when the CID is known and trusted. Furthermore, the IPFS will enable KR to break the stored contents into chunks of blocks consisting of **256*1024(262144)** bytes. Therefore, if the size of the uploaded content is **262144**, two CIDS will be generated, one of which will refer to the first block of **262144** bytes and the other will refer to the last block (Tindall 2016).

7.2.9. INTERPLANETARY NAMING SYSTEM (IPNS)

The IPNS is a mechanism that allows entities to generate meaningful names of the stored contents. Generally, the stored contents are identified by the unique hashes; however, such an approach can be often harder when sharing or accessing a content. Therefore, meaningful human-readable content names are required for efficiency. The IPNS enables entities to comprise a mutable, human-readable name referencing a particular IPFS content address. When contents are published to IPNS, it points to IPFS hash. The record gets signed by users' public key and thus confirming authenticity and ensuring no alteration without proper consent.

7.2.10. BITSWAP PROTOCOL

Bitswap serves as the core component of IPFS simplifying the knowledge content exchange among the KR network peers. It is a message-based protocol enabling all messages to consist of want-lists or blocks. In a Bitswap protocol, the distributed nodes share a want list with the associated peers and consist of the list of CIDs for the blocks they aim to obtain. Each node keeps track of the blocks its peers expect from them. Once a block is received, the node confirms its list to determine whether any connected peers desired that particular block. The major functionalities of Bitswap are; (i) obtaining the blocks asked by the client and (ii) transmitting blocks in its authority to peer nodes that require them. In contrast, a BitTorrent exchange comprises data to be exchanged within a single torrent.

7.2.10.1. VERSIONING AND HISTORY

The KR will enable the knowledge-related content to be versioned and thus several versions resulted in from a content can exist in the network. Each version will comprise a CID and can be accessed by referencing the particular version. Versioning can be advantageous for applications requiring immutability and allow them to track the changes. In addition, KR will consist of a historical record of all the changes made to the stored knowledge contents and thus when new content is added, it will enable KR to contain all the previous versions generating a historical lineage of resources. The history of the contents is inherent in the Merkle DAG structure (Steichen 2018).

Figure 64 shows how the versioning of contents works and refer to the previous version.

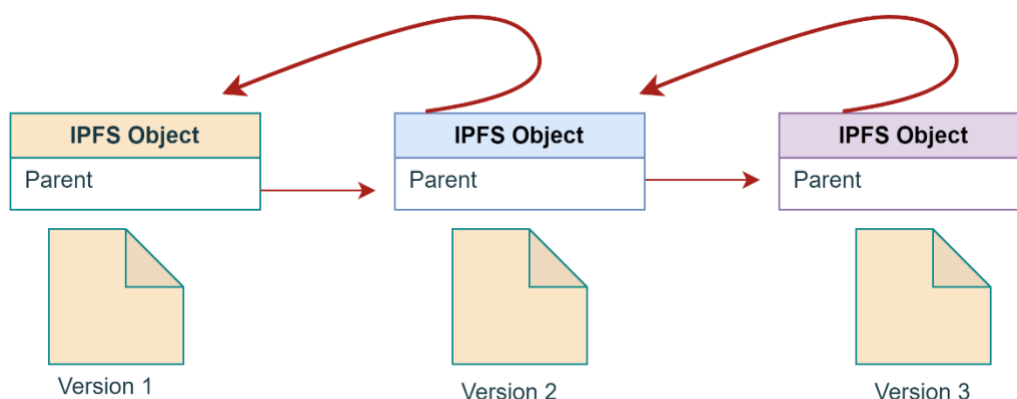


Figure 64 Versioning method of KR referring to previous version of the knowledge content.

7.2.10.2. CRYPTOGRAPHIC SECURITY

IPFS comprises several cryptographic security methods in place to ensure complete content security, privacy, and secure communication in the Network. Below are some of the key approaches.

- **Content Integrity:** The cryptographic hash function generates a unique CID which will ensure the integrity of the content, making it tamper-resistant. Thus, resources obtained from IPFS will remain unchanged.
- **Encryption:** IPFS comprises end-to-end encryption for secure transmission of contents. Moreover, resources can also be encrypted before being added to the repository thus ensures that only the authorised parties with decryption keys can access it.
- **Secure Communication:** The employed encryption mechanism for secure communications prevents various attacks, such as man-in-the-middle-attacks, eavesdropping, etc., and thus enhances the content transmission secure and confidential.
- **Digital Signatures:** For content authentication and verification, IPFS supports digital signatures which ensures that the content has come from a trusted source.
- **Peer Identity:** IPFS implements cryptographic key pairs to form the peer identity. This ensures secure peer-to-peer communication among the peer nodes confirming that the content is shared with the trusted network entity.

7.2.10.3. AUTOMATED ANALYSIS

Automated analysis tools and techniques have emerged as crucial assets in extracting meaningful insights, patterns, and knowledge from these vast troves of data. As the reliance on data continues to grow, the need for effective and efficient data analysis tools becomes more critical. Figure 65 depicts the process of automated analysis utilizing panda profiling.

Python, an open-source, high-level programming language, frequently emerges as the preferred choice for automating data analysis tasks. Python offers a rich array of specialized libraries tailored for data, analysis, and visualization.

For the Automated Analysis in the scope of TRUSTEE the chosen technology is python's **pandas-profiling** library which is an open-source Python library that automates and simplifies the initial data exploration and profiling

process. It offers a convenient way to generate comprehensive, insightful reports about datasets, providing a quick overview of its characteristics and potential issues.

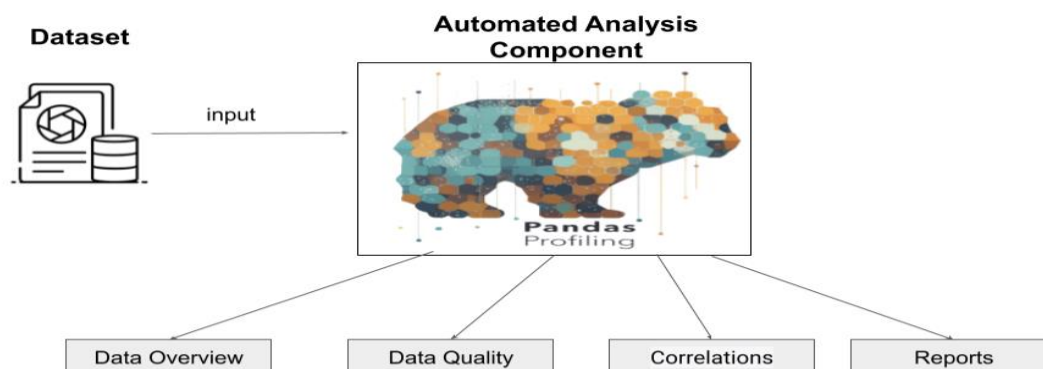


Figure 65 Automated Analysis using pandas profiling.

7.2.10.4. KNOWLEDGE GRAPH

The Knowledge Graph within the TRUSTEE Knowledge Repository benefits extend beyond mere data organization; it enriches the user experience by simplifying data discovery, providing context, encouraging interdisciplinary exploration, streamlining data integration, ensuring data quality, and promoting community engagement. These advantages collectively will empower our users to make more informed decisions, drive innovation, and extract greater value from the Knowledge Repository. The integration of the Knowledge Graph can enhance the structural representation of the knowledge stored within the knowledge repository.

7.3. REQUIREMENTS

In this chapter, the compliance of the KR ABB of TRUSTEE is investigated with regards to the user-centred, technical, legal, and socio-ethical requirements that were reported in D2.1 (TRUSTEE consortium, 2023). The results of this investigation provide further preliminary insight regarding with the aforementioned requirements addressed during the development and by the implementation of the KR and are presented in the following sub-sections. It is worth noting that only the Mandatory, Desirable, and Optional requirements have been considered at this point. Furthermore, only the currently compliant or partially compliant requirements have been considered. As the implementation of the KR progresses during the development of the TRUSTEE Platform and the maturity of the ABB increases, the investigation of the compliance will evolve and further updates and/or enhancements will be provided in upcoming versions of this deliverable, namely D3.2 and D3.3.

User-centred requirements were identified through the UX Research that was conducted under T2.3. These requirements were reported in D2.1 (TRUSTEE consortium, 2023), while they were also utilized in order to extract technical, functional and non-functional requirements, that were also outlined in D2.1. Table 17 below presents the interplay between user-centred requirements and technical requirements, that is the functional and non-functional requirements that address each user-centred requirement. The user-centred requirement ID is used to reference the respective requirement reported in D2.1 (TRUSTEE consortium, 2023) while the requirement's description is also provided alongside whether the KR is currently fully or partially compliant with the user-centred requirement, i.e.,

indicating whether technical requirements have been identified in order to address the respective user-centred requirement. The ID of the technical requirement, as reported in D2.1 (TRUSTEE consortium, 2023), that is addressing each user-centred requirement is also provided in the last column of the table. Additionally, it is possible than more than one technical requirement addresses a single user-centred requirement.

Table 17: KR – User-centred requirements

Req id (WP2)	Req. Description	Compliance	How is addressed
USR-015	TRUSTEE should be an open-accessed data repository	Yes	Accessible to authorised entities.
USR-052	TRUSTEE should extract knowledge without complex analysis methods and tools	Yes	No additional tools or methods required to extract the stored knowledge from KR. The CID identifies each knowledge and works as a key to obtain particular knowledge.
USR-030	TRUSTEE should include multiple querying functions	Partially	The knowledge stored in KR can be searched through their CID (Content Identifier) to obtain the original resource.

Table 18 presents the preliminary results of the investigation regarding the compliance of the KR with the technical, functional and non-functional requirements specified in D2.1. The ID of the respective technical requirement is used as a reference to D2.1, while the requirement's description is also provided alongside whether DA complies with the requirement and how. Further updates and/or enhancements regarding the approach presented below for addressing each technical requirement will be provided in upcoming versions of this deliverable, namely D3.2 and D3.3.

Table 18: KR – Functional and non-functional technical requirements

Req id (WP2)	Req. Description	Compliance	How is addressed
Req-KR-FUNC-1	The Knowledge repository will be based on IPFS network which will consist several decentralised nodes. The nodes will be the building blocks of the network and play a crucial role as they will be responsible for storing all datasets in the peer-to-peer network.	Yes	KR will enable to store knowledge within the distributed nodes.
Req-KR-FUNC-2	A monitoring tool will be configured for integration with the knowledge repository. The tool should enable to obtain local node information, content details, content	Yes	KR should allow to gain an insight from the distributed network to determine the network performance,

	sharing information, peers, information, content search functionalities, content transfer history, and so on. ledger.		knowledge-based resource distribution, etc.
Req-KR-FUNC-3	IPFS cluster is a distributed application which will be integrated with the Knowledge Repository (IPFS Network). It will work as a sidecar to IPFS peers, maintaining a global cluster pinset and allocating its items to IPFS peers.	Yes	IPFS cluster will enable to comprise a cluster of nodes to maintain high availability of the knowledge resource through proper resource management.
Req-KR-FUNC-4	IPFS network does not provide automated analysis functionalities. Thus , other tools and platforms should be integrated to utilise the automated analysis functionality for the stored datasets in IPFS. Data stored in the knowledge repo (IPFS network) might be accessed by Apache Spark, etc.	Yes	Automated analysis will automate and simplify the knowledge resource exploration and profiling.
Req-KR-FUNC-5	IPFS does not provide knowledge graph functionalities. Thus, other tools and platforms should be integrated to utilise the Knowledge Graph functionality for the stored datasets. The knowledge graph should aggregate the information about the structure of the data itself.	Yes	Knowledge Graph technologies will be integrated with the Knowledge Repository, allowing seamless access and utilization of the Knowledge Graph's capabilities for data stored in IPFS. This integration will enable the Knowledge Graph to aggregate structural information about the data while leveraging IPFS for data storage and retrieval.

A preliminary analysis of the compliance with the legal and socio-ethical requirements defined in the D2.1 (TRUSTEE consortium, 2023) is shown in ANNEX 2 and ANNEX 3.

7.4. ARCHITECTURE

This section showcases both the high level and low-level architecture of KR along with the high-level overview of knowledge graph. Figure 66 defines the phases of knowledge storage in KR. A user stores content to the KR cluster and obtains CID in response. The CID is then recorded to the ATR which comprises 2 types of access control. While the CIDs recorded to the private state of ATR, only the authorised entity access to that private ledger will be able to retrieve it, recording the CID's to the blockchain state will ensure the immutability of the stored CIDs and can be accessed by all authorised parties.

7.4.1. HIGH LEVEL

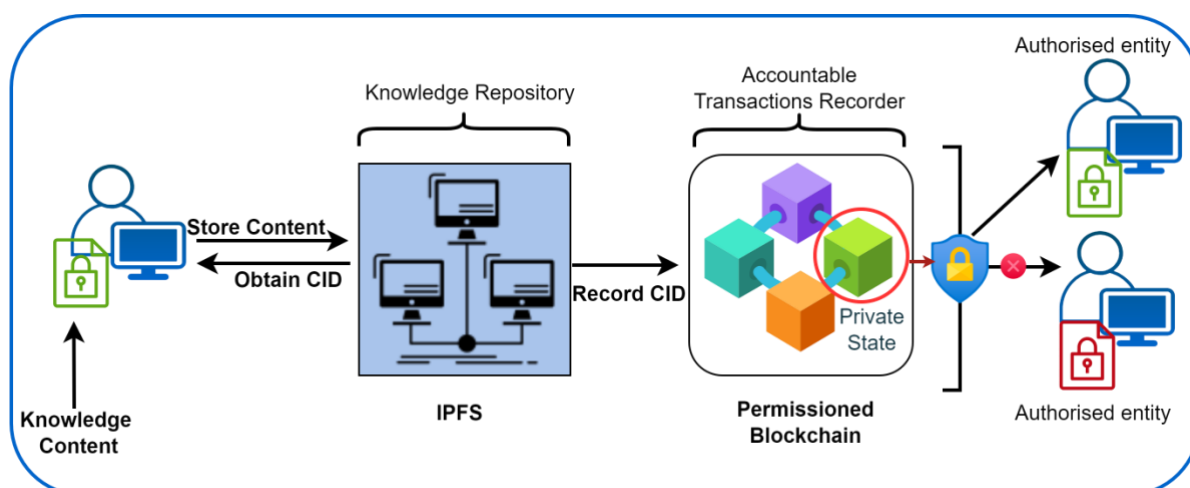


Figure 66 High level architecture of KR enabling user to store content to the repository.

Knowledge Graph:

Designing and developing a Knowledge Graph requires a careful selection of technologies to ensure its efficiency, scalability, and adherence to our project's goals, including the FAIR principles.

In this section, we provide a high-level overview of the key technologies that can be employed in the creation of our Knowledge Graph and which will be presented in detail in the next versions of this live document.

Graph Database Management Systems (DBMS): Graph databases such as GraphDB, Amazon Neptune, or Virtuoso are at the core of Knowledge Graph development. These DBMSs are designed specifically to handle graph structures efficiently, allowing for the storage and querying of interconnected data entities. They enable us to model and store data in a way that reflects real-world relationships, making it easier to navigate and analyze complex datasets. The TRUSTEE Data Lake already provides the GraphDB store for the storage of all semantically annotated datasets concepts that can be accessed from the Knowledge Graph and Knowledge Repository upon request.

Semantic Web Technologies: Semantic web standards like RDF (Resource Description Framework) and OWL (Web Ontology Language) are essential for representing and encoding the semantics of data. The Knowledge Graph will have access to this semantic data from the TRUSTEE Data Lake, ensuring semantic interoperability.

Data Integration and ETL Tools: To aggregate data from various sources and transform it into a format suitable for the Knowledge Graph, Extract, Transform, Load (ETL) tools like Apache Nifi, Talend, or Apache Spark can be employed. These tools streamline the process of data ingestion, cleansing, and transformation, ensuring data quality and consistency.

Knowledge Representation Models: Techniques like Word2Vec, FastText, or BERT can be employed to embed textual data into vector representations. These embeddings enhance the Knowledge Graph's ability to understand and semantically link text-based information, bridging the gap between structured and unstructured data.

APIs and Web Services: Integration with external data sources and the rest of the TRUSTEE ABBs will be achieved through APIs and web services. Technologies like RESTful APIs, GraphQL, or IPFS APIs will be used to facilitate real-time data exchange, allowing our Knowledge Graph to stay up-to-date with any data changes.

Data Visualization Tools: To make the Knowledge Graph accessible and comprehensible to users, data visualization libraries like D3.js, Plotly, or NetworkX can be employed. These tools enable the creation of interactive and informative visualizations that aid in data exploration.

Figure 67 presents an overview of the proposed technologies that can be useful for the design and implementation of the Knowledge Graph.

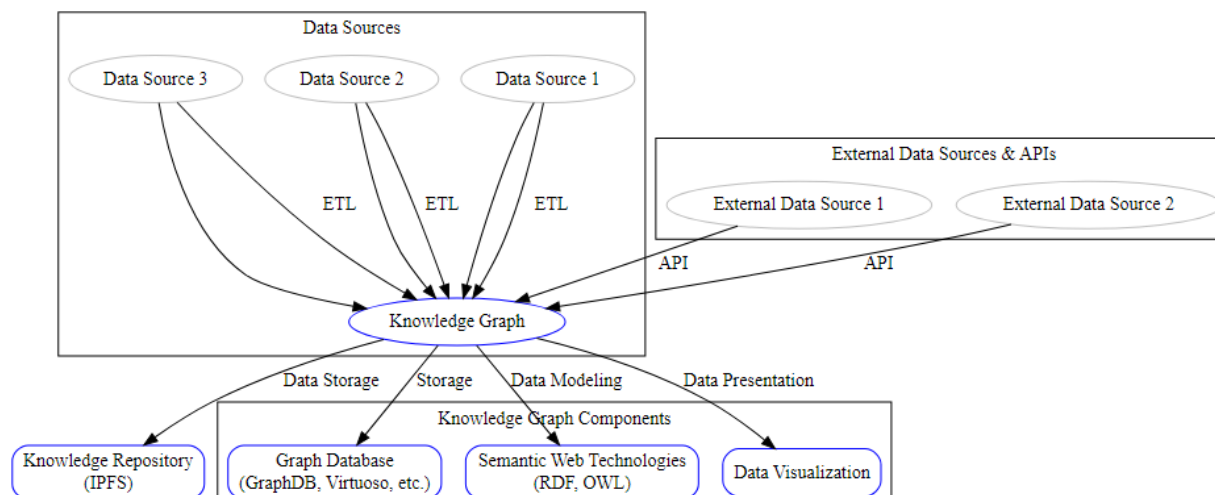


Figure 67 High level overview of Knowledge Graph technologies

The successful design and development of our Knowledge Graph will rely on a well-thought-out combination of the aforementioned technologies. By leveraging graph databases, semantic web standards, data integration tools and other relevant technologies, we can create a robust and FAIR-compliant Knowledge Graph that serves as a valuable asset within the TRUSTEE Knowledge Repository.

7.4.2. LOW LEVEL

Figure 68 shows the functioning of KR implemented using peer-to-peer file sharing protocol. Four peer nodes are part of the distributed network where each of them comprises a distributed hash table (DHT). The DHT provides information about the content and the peer that holds it. The DHT table is updated whenever a peer requests a content from other peer nodes showing that the content has been pinned to that node. The final phase shows the storage of the CID to the blockchain for enhanced security and traceability.

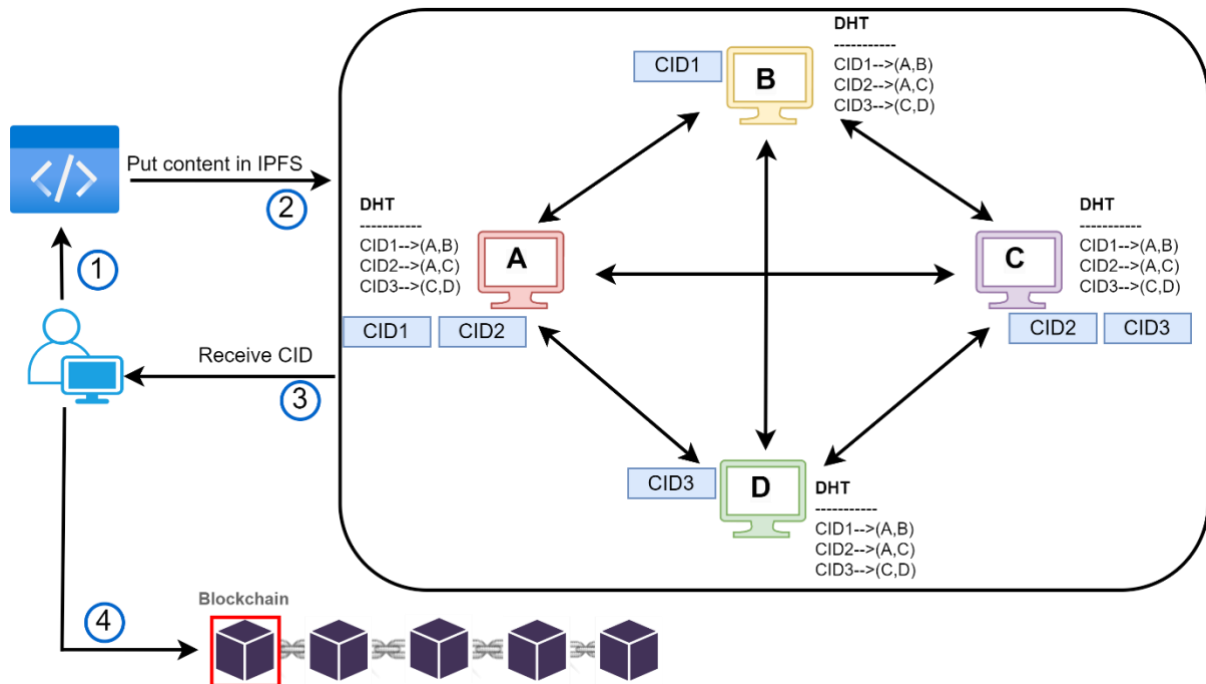


Figure 68 Inner functioning of KR

7.5. IMPLEMENTATION

During the first implementation phase, both the private and public IPFS have been considered for KR. The implementation of both has helped determine the advantages and disadvantages and choose the suitable option. A private Knowledge repository has been implemented which allows to storage of knowledge-related content, pin contents that hold significance, and unpin contents that need to be removed through garbage collection. Additionally, the implementation also involves outside communication through IPFS API enabling to storage of knowledge-related content and generating individual CIDs for each domain so that only the resource that is required can be extracted without requesting the entire content. Furthermore, the integration of KR with the IPFS desktop shows how contents can also be uploaded through the Desktop and all the peers a node is connected with. In the next implementation phase, the integration with ATR will be considered, along with this the implementation of automated analysis and knowledge graph functionalities will also be considered. Figure 69 shows the implementation of KR enabling TRUSTEE ABBs to interact utilising the IPFS API to store knowledge.

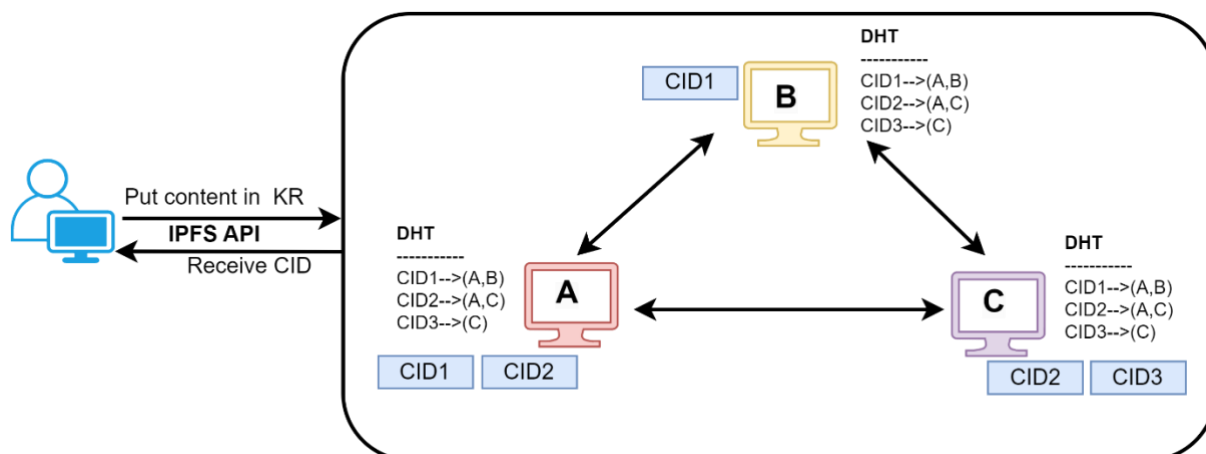


Figure 69 TRUSTEE KR allowing user to interact to store knowledge.

7.5.1. FUNCTIONAL DESCRIPTION

The description of the functionalities of KR that have been implemented for the establishment of KR are as follows;

- **IPFS Cluster:** IPFS cluster enables the inclusion of additional features to KR for knowledge-based content replication, pinning, and coordination among groups of IPFS nodes. The cluster is designed in a way that allows it to operate a cluster of nodes ensuring full content availability, redundancy, and fault tolerance. IPFS cluster is useful to enhance the knowledge-based content-sharing performance and data management in a distributed environment. Furthermore, the cluster nodes maintain a global pin set which they have complete control over. It is important to note that when utilising the IPFS cluster the replication of contents is not designed to be decentralised.

The main advantages of the IPFS cluster are as follows:

- **Content pinning and high availability:** The cluster ensures high availability and content redundancy. The pinning preserves specific content to all cluster nodes, even while some nodes become unreachable, the pinned contents can still be accessed via other nodes within the network.
- **Dynamic Cluster Management:** The cluster allows to addition or remove of nodes from the cluster enabling a flexible approach to managing the cluster of nodes.

```
> ipfs - cluster -ctl pin add QmbJBY6JtyDjFE9KtDrkq446DVxK7SLn1CmWx2iUvKz6g
QmbJBY6JtyDjFE9KtDrkq446DVxK7SLn1CmWx2iUvKz6g :
> ipfs - cluster -0 : PINNED | 2023 -06 -09 T09 :36:28 Z | Attempts : 0 | Priority : false
```

Listing 5 An example of pinning a content utilising IPFS cluster

- **IPFS API:** The IPFS API comprises a set of endpoints enabling users to interact with the IPFS network. It provides an efficient method to access the KR functionalities which can include storing, retrieving, or managing knowledge resources through HTTP requests. The IPFS API offers various functionalities which include; (i) adding content to the KR by making API calls. (ii) retrieving contents using CID ensuring secure and reliable content access. (iii) interacting with Merkle DAG by adding, getting, and resolving DAG nodes. (iv) interact with the KR resource through HTTP gateway enabling access to the content

through standard web browsers. (iv) flexible object management by creating, deleting, and updating KR content.

IPFS Desktop: The IPFS desktop is a user-friendly web-UI tool which combines IPFS node, file manager, peer manager and content explorer into a single package. The desktop serves the entities to communicate with the IPFS network efficiently. It establishes a powerful bridge with the Knowledge Repository (KR) enabling entities to upload and access knowledge content. The desktop provides taskbar shortcuts to expedite workflows. Regardless of whether a node exists or not within the IPFS network, the IPFS desktop can allow to configuration of nodes based on the requirements. It functions as a control panel and file browser performing automatic updates ensuring the latest functionalities are achieved while connected to the network.

The main functionalities of IPFS Desktop are;

- **Knowledge Management:** The IPFS desktop includes a file manager which enables entities to upload, download and organise knowledge-related content in the IPFS storage.
- **Node Management:** The IPFS desktop enables entities to run and manage all the IPFS nodes without going through any complex node management phases.
- **Peer Management:** The peer manager allows for monitoring and management of the p2p connections with other nodes within the IPFS network. It also enables one to have full control over the network settings and view network statistics.
- **Content Explorer:** The content explorer can play a significant part by allowing users to navigate and explore the Knowledge Repository to discover, view and access all knowledge content.
- **Streamlined Integration:** The IPFS desktop allows integration with various IPFS functionalities in a single application and thus simplifies the user experience.
- **Knowledge Pinning:** The pinning refers to the method which specifies that the stored resource be retained and maintained in the network permanently. The pinning functionality ensures that any knowledge content that is stored in the Knowledge Repository can be accessed indefinitely and will not be removed while the Garbage Collection is executed. The pinning plays a crucial ensuring content persistence and availability.

There are three kinds of pinning that can be exercised;

- **Direct Pins:** This refers to pinning an individual block without having any other blocks associated with it.
 - **Recursive Pins:** It refers to pinning specific blocks along with all of its child blocks.
 - **Indirect Pins:** Indirect pins are required when the parent block of a given block is recursively pinned.
- **Garbage Collection:** The Garbage collection mechanism ensures that all the stored resources to KR are maintained efficiently and when a resource is no longer needed, it can be unpinned for garbage collection to remove the unreferenced content. The garbage collection is crucial as it helps managing the storage efficiently and ensure that the KR storage space is fully optimised. By executing the garbage collection feature, the KR will be able to maintain more organised and streamlined storage settings and enhance its performance.
- **Blockchain Integration:** Integration of Knowledge Repository with Blockchain provides several advantages. The Accountable Transactions Recorder (ATR) will be implemented utilising permissioned blockchain. Since blockchain is not very efficient in storing large contents, integration with KR can reduce the storage burden. Also, it provides an added layer of security by recording the transactions in the blockchain. The integration enables to improve the overall performance while ensuring integrity and

privacy. In addition to that, the fast and efficient content retrieval method of KR will enable to perform operations at a fast pace.

- **Knowledge Graph:** The Knowledge Graph, as an important part of the Knowledge Repository ABB, serves as a pivotal tool for organizing and harnessing information effectively. It will fulfil three fundamental functionalities and purposes, each contributing to the overall success of our project:
 - **Structural Information Aggregation:** At its core, the Knowledge Graph excels in collecting and representing structural information about the data within our repository. It acts as a dynamic map, cataloging the relationships and hierarchies that exist among various pieces of data. By doing so, it enables us to gain a comprehensive understanding of the underlying data structure. This structural knowledge empowers researchers and stakeholders to navigate, access, and analyse the data with greater ease and precision. It essentially boosts operations such as data integration and interoperability, ensuring that data from different sources can be harmonized and utilized cohesively (Sachdeva & Bhalla, 2022).
 - **Relational Information from Multiple Sources:** Another critical function of the Knowledge Graph is to capture and encode relational information originating from diverse data sources. By assimilating these relationships, whether they are semantic, thematic, or contextual, the graph enhances our ability to draw meaningful insights and make informed decisions. This cross-referencing of information across various domains and sources is essential in TRUSTEE because of the multidisciplinary nature of the datasets and can thus uncover hidden patterns, identify knowledge gaps, and facilitate integrative research (Elshani et al., 2022).
 - **Supporting FAIR Principles:** Adhering to the FAIR (Findable, Accessible, Interoperable, and Reusable) principles is paramount in modern data management and one of the important objectives of the TRUSTEE ecosystem. The Knowledge Graph aligns perfectly with these principles, contributing significantly to their realization. It enhances data findability by providing a structured and organized framework for efficient search and retrieval. Its role in ensuring data interoperability is indispensable, as it bridges the semantic gaps between disparate datasets. Moreover, the graph's ability to capture and represent data provenance and metadata makes data more accessible and reusable. By embodying the FAIR principles, our Knowledge Graph not only optimizes data management but also promotes data sharing and collaboration within and beyond our research project, fostering a culture of openness and transparency (Pestryakova et al., 2022).
- **Automated Analysis:** TRUSTEE's Automated Analysis supports the following features:
 - **Knowledge Overview:** summary report that includes basic statistics like mean, median, standard deviation, and more for each numerical column
 - **Quality Assessment of Knowledge Resource:** identification of missing values, duplicates, and visualizations like bar charts and heatmaps to highlight correlations between variables
 - **Knowledge Distribution Visualizations:** histograms, box plots, and scatter plots to visualize the distribution and relationships between data points
 - **Correlation Analysis:** correlation matrices, making it easy to identify relationships between numerical variables
 - **Interactive Reports:** interactive HTML files that allows data exploration the data visually

Table 19 The core functionalities of first implementation

Functionality	Description	Functional

IPFS Cluster	A cluster of 3 KR nodes which can store TRUSTEE knowledge contents.	Yes
IPFS Desktop	A GUI application which provides efficient communication with KR	Yes
IPFS API	IPFS API enabling applications to interact with KR for resource storage, retrieval, etc.	Yes
IPFS Pinning	Pinning ensure the knowledge resource is available and cannot be removed unless unpinned from all nodes.	Yes
Garbage Collection	Garbage collection removes unpinned content	Yes

7.5.2. OPERATIONAL DESCRIPTION

Figure 70 and Figure 71 shows the sequence of knowledge storage to the KR and recording of the CID to ATR. The Dashboard integration enables users to view the knowledge related content.

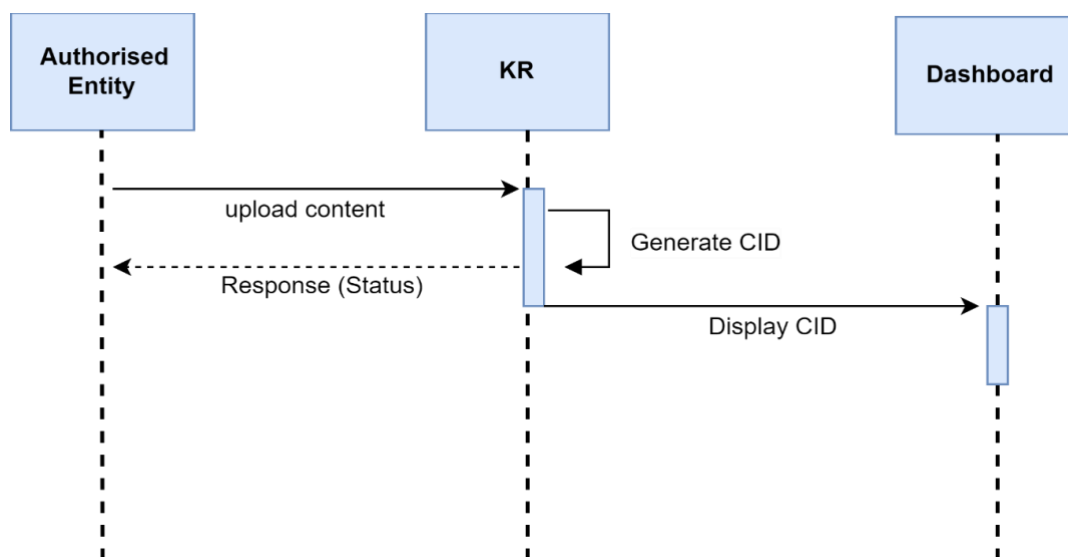


Figure 70 The interaction to KR for Knowledge storage by authorised TRUSTEE entity.

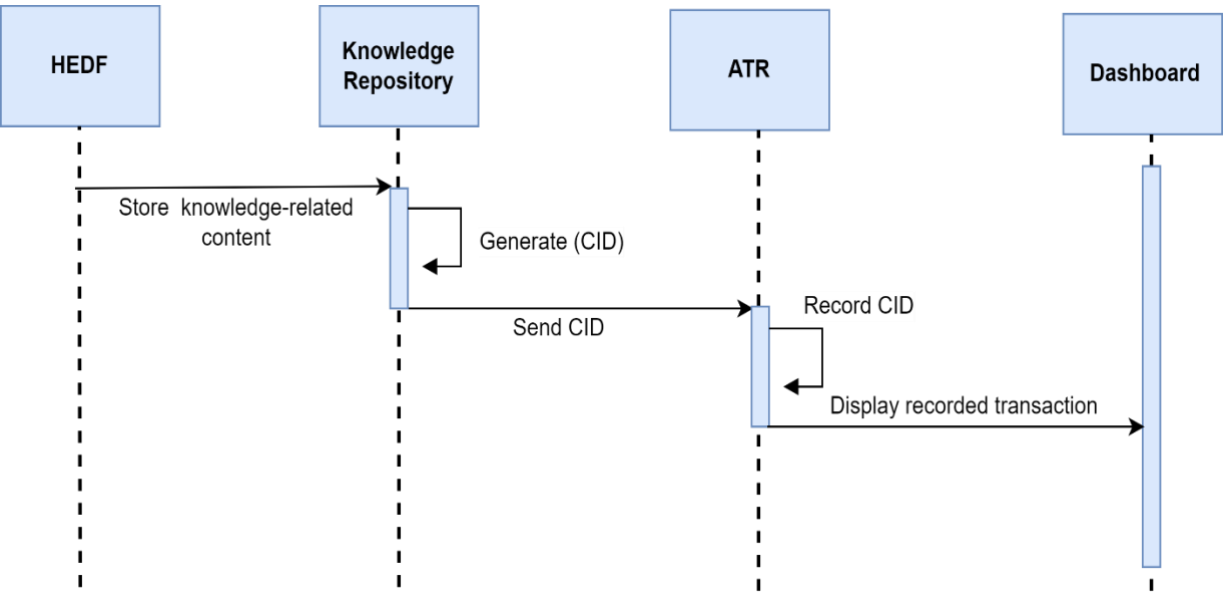


Figure 71 Interaction from an ABB for the storage of knowledge to KR and the recording of CID to ATR.

7.5.3. TECHNICAL DESCRIPTION

In this section, a comprehensive overview of the project's implementation details, covering Input/Output/Data Models, Graphical Interface (including mock-ups and frontend components), and the API will be presented. It highlights the technology stack, programming languages, and development methodologies utilized in the project's construction, along with data flow, user interactions, and data validation mechanisms.

7.5.4. TECHNICAL DETAILS

The Technical Details section provides an in-depth overview of the development, installation, and deployment aspects of the TRUSTEE platform. This section is crucial for technical stakeholders and development teams to understand the tools, technologies, and infrastructure required to successfully implement and deploy the KR ABB. Table 20 shows the technical details required for the first phase implementation of the KR and Table 21 shows the generic structure which has been considered for communication in order to store knowledge to KR utilising IPFS API. The KR responds with 4 CIDs, one for the entire content and three for particular domains enabling retrieval of either the whole content or partial content according to the needs.

Table 20 Technical details of first implementation phase

DEVELOPMENT	
Type (Software/Hardware)	Software
Operating Systems	Linux (Ubuntu), Windows
Database and databases tools	Local cache, dedicated storage defined by the node/user.

GUI	Command-line driven
Programming language	n/a
Development Environment	n/a
INSTALLATION AND DEPLOYMENT	
Software Requirements	IPFS
Hardware Requirements	Standard hardware requirements
Containerization	Optional
Communications	-

Table 21 Structure for knowledge storing to KR

<pre>{ "content_data": { "content1": { "process_name": " My Process-01-AUTO ", "type": "search_query_results", "value": "<content_data>" }, "content2": { "process_name": " My Process-01-AUTO ", "type": "fusion_results", "value": [1234, 5678] } }, }</pre>

```

"content3": {
  "process_name": " My Process-01-AUTO ",
  "type": "computation_results",
  "value": "<content_data>"
}
}
}

```

7.6. VALIDATION

- Figure 72 shows the storage of knowledge-related contents to the KR from an outside application interacting through IPFS API. The KR responses with 4 CIDs, one of which can be used to obtain the entire content while the rest 3 CIDs can be used to retrieve content 1, 2 and 3 respectively.

```

sarwar@Test:~/sarwar_test.git/trustee/ipfs_main_impl$ node dapp.js
CID of full knowledge content: QmRJ7i4LS4MnJbGfExBbZwqq2is11tPhT87TLVuN7ERis
CID of Health related knowledge: QmUkoSpbCcPrtCrVLWA8fBjwpJjcSxJx7ixMNwFh7rDRn5
CID of Auto related knowledge: QmYV2pJWbQ5CYJYR3yFSc6W6FS7yTXzLfbggvNAY8kt9U4
CID of Space related knowledge: QmTKE2s2ZWboDUEzRLpvXedVvdvY5Wefd1Fe75RiavwFYa
sarwar@Test:~/sarwar_test.git/trustee/ipfs_main_impl$

```

Figure 72 Content stored to KR.

- Figure 73 shows the retrieval of the full knowledge content from KR.

```

sarwar@Test:~/sarwar_test.git/trustee/ipfs_main_impl$ node dapp.js
CID of full knowledge content: QmRJ7i4LS4MnJbGfExBbZwqq2is11tPhT87TLVuN7ERis
CID of Health related knowledge: QmUkoSpbCcPrtCrVLWA8fBjwpJjcSxJx7ixMNwFh7rDRn5
CID of Auto related knowledge: QmYV2pJWbQ5CYJYR3yFSc6W6FS7yTXzLfbggvNAY8kt9U4
CID of Space related knowledge: QmTKE2s2ZWboDUEzRLpvXedVvdvY5Wefd1Fe75RiavwFYa
sarwar@Test:~/sarwar_test.git/trustee/ipfs_main_impl$ ipfs cat QmRJ7i4LS4MnJbGfExBbZwqq2is11tPhT87TLVuN7ERis
{"content_data":{"content1":{"process_name":"My Process-01-Health","type":"search query results","value":"content_data"},"content2":{"process_name":"My Process-01-AUTO","type":"fusion_results","value":"12345678"},"content3":{"process_name":"My Process-01-Space","type":"computation_results","value":"content_data"}}}
sarwar@Test:~/sarwar_test.git/trustee/ipfs_main_impl$

```

Figure 73 Interacting with KR to retrieve whole content.

- Figure 74 shows the retrieval of selected content from KR that are based on particular domain.


```

sarwar@Test:~/sarwar_test.git/trustee/ipfs_main_impl$ node dapp.js
CID of full knowledge content: QmRJ7i4LS4MnJbGfExbBbZwqq2is11tPhT87TLVuN7ERis
CID of Health related knowledge: QmUkoSpbCcPRTCrVLWA8fBjwpJjcSxJx7ixMNwFh7rDRn5
CID of Auto related knowledge: QmYV2pJWbQ5CYJR3yFSc6W6FS7yTXzLfbggvNAY8kt9U4
CID of Space related knowledge: QmTKE2s2ZWboDUEzRLpvXedVvdvY5Wefd1Fe75RiavwFYa
sarwar@Test:~/sarwar_test.git/trustee/ipfs_main_impl$ ipfs cat QmUkoSpbCcPRTCrVLWA8fBjwpJjcS
xJx7ixMNwFh7rDRn5
{"process_name":"My Process-01-Health","type":"search_query_results","value":"content_data"}
sarwar@Test:~/sarwar_test.git/trustee/ipfs_main_impl$ ipfs cat QmYV2pJWbQ5CYJR3yFSc6W6FS7yT
XzLfbggvNAY8kt9U4
{"process_name":"My Process-01-AUTO","type":"fusion_results","value":"12345678"}sarwar@Test:
sarwar@Test:~/sarwar_test.git/trustee/ipfs_main_impl$ KE2s2ZWboQmTKE2s2ZWboDUEzRLpvXedVvdvY5
Wefd1Fe75RiavwFYa
{"process_name":"My Process-01-Space","type":"computation_results","value":"content_data"}sa
rwar@Test:~/sarwar_test.git/trustee/ipfs_main_impl$

```

Figure 74 Partial retrieval of content.

-
- Figure 75 shows the lists of pinned contents which remain in the repository until they have been removed from all KR nodes.

```

sarwar@Test:~/sarwar_test.git/trustee/ipfs_main_impl$ ipfs pin ls
QmZaEQdUN9aj2UXzac2QHgDq5k2PHm6kMi17zJpRwwgKnm recursive
QmZabhuxsq4JJUyPh5dYmWxGduGrKNVxj9mtm28AxZqhuc recursive
QmR4rV5CzZ1vmugCMtCTMGDqb9aPbiKJvZDEo19fYQR9U recursive
QmRY6a6jgUyQfLR2cm7kuso7k2kFGBbJkYqLZZ3U19f4e recursive
QmSxwd6nCD43UjUqUWfufb59rSsDEd8cNvfzA4112Kq1 recursive
QmUNLLsPACcz1vLxQVxKqLX5R1X345qqfHbsf67hVA3Nn recursive
QmUJW6Q2VzjHEVxrWmvdZggVLnn3eDcGGtX5EMk67EJy93 recursive
QmZSajYUSQbB1a2DcbVA6Dyx48gZMUeWQDFWziwkVEQthT recursive
QmZsZkzhfxD5o649xr69Zp5MER7AuSXBtpXxyuS15fywg recursive
Qmew4Z7XCxhDiHJRQgDrq21BHDJb6uZKQGapQ7aLk6Cias recursive
QmRGegghM2skT7GMZR48VArv7aPqtqVv1efCSm3n7eNy7u recursive
QmRgPVzsaCWW8WX2FnqQgER8X21qCqFwQSDTbkFhpg7q recursive
QmSfcQ7ugu8adtGeKkm1VKL2zyEvNH64iHD3hg9oFn9F78 recursive
QmTKE2s2ZWboDUEzRLpvXedVvdvY5Wefd1Fe75RiavwFYa recursive
QmYV2pJWbQ5CYJR3yFSc6W6FS7yTXzLfbggvNAY8kt9U4 recursive
QmYnjKEkDPrA6kpcnKdcPg5UkZ2PodMF6CDS7Bhw3BZxm recursive
QmZv1JRh5XxpKVd2236cmQ21e6BcfQcY5wzyKVnrg1X4ys recursive
QmfVDjrEXXhSc96CytWA4eTHCRRSsDYJNq7AM6HjP2hShx recursive
QmcmV74nBCahd2V7cmqun2j4syCdPgEESUnF5sBP4Fc2fv recursive
Qmcw98tFQ7Syjhwgqom6JkAqTfzGxzQGzXSCqYdG42PM recursive
QmeCu31Qo7J51h83gxDSq8VJLh9hQys3j6vF6k4lB3KprC recursive
QmRJ7i4LS4MnJbGfExbBbZwqq2is11tPhT87TLVuN7ERis recursive
QmRYfujVzWdx7H1RtL6SgBLw7yF583cPuqa8MsaantzKLF recursive
QmUGVtunZnCGgve7RGxPWL7c894EoBNUkcBXXSRgES1E7X recursive
QmUkoSpbCcPRTCrVLWA8fBjwpJjcSxJx7ixMNwFh7rDRn5 recursive
sarwar@Test:~/sarwar_test.git/trustee/ipfs_main_impl$ ipfs pin rm QmZaEQdUN9aj2UXzac2QHgDq5k2PHm6kMi17zJpRwwgKnm
unpinned QmZaEQdUN9aj2UXzac2QHgDq5k2PHm6kMi17zJpRwwgKnm
sarwar@Test:~/sarwar_test.git/trustee/ipfs_main_impl$

```

Figure 75 Pinned content and removal of pinning for execution of garbage collection.

- Figure 76 shows the list of unpinned objects and the execution of GC for deletion of the unpinned content.

```

sarwar@Test: ~/sarwar_test.git/trustee/lpfs_main_lmpl$ lpfs repo stat --human
NumObjects: 1134
RepoSize: 11 MB
StorageMax: 10 GB
RepoPath: /home/sarwar/.lpfs
Version: fs-repo@14
sarwar@Test: ~/sarwar_test.git/trustee/lpfs_main_lmpl$ lpfs repo gc
removed bafkrethcushudcqrj71cbltwx27ps560la2ordnybg7t5lgtmv5epog5p4
removed bafkrethskvby4pyrjnp1dsbft62t43qqj7jh7ya4yeyq7f3zhyfxpnp4
removed bafkretday3dhkasdraxclqlvgv2hmo54r7yl54m2akapmfhhak4evnx4
removed bafkreifu4vfwkrrdgmuloovolsjjiif67zxbul6nrmfp4bdaex4qr7l3r5xu
removed bafkreic27ssqtdnrwecfz3f3sffn52s1jffxeegjwvsnoy6323h3wzga
removed bafkreig4ftwlgstlvmf5tlanhygzyivrcjesn50pe4uciabwzs2o27xy
removed bafkreic6fn3dntrof6aceltuv44wsray6pmvuj7hys7eumv3j5uvxbgza
removed bafkreifgmbtihtkvtwhykjj2cvpa3lvvdpmc65dcr6egj33q4mx2wahhgzi
removed bafkreiflduhsurxu6kt5vxs2apx6tnhd3wclg375ptsen4wtznch72svy
removed bafkreidgdqcdnn4mav2tgcpi7etgbyzdklp34uvn5odghgvhschtvsq
removed bafkreiflh2v5omhse4lmmrbw373kxtur2tp7mchsxu555pj3bbxqoaf4
removed bafkreicvvg4n4i3tnopltkmrwtidoqz24gsjvif3dgszlosd76pdlbqf4
removed bafkreigndylek436cvcu7lrmhnpmtaaxqfz7stcdkycok5wxtoppfqf4
removed bafkreifm5hcuae4tje4scxcqcaoz2yue24mpvbcx5en12lml6gxnbpwpi
removed bafkreitbv7gsgkx45lclpyys12wqgywnl2h27mpzmkyluo6psrx3wj1
removed bafkreig4lr1lukevnyu5n6n77c5ew7z2drtyc6dnzeuhea3qvw3fposjtrq
removed bafkreid2ayxufttjgrgvro2lmp7ugx2p3jvkk3c3jystcoppxzoqmrq
removed bafkreiteqjdhyo5zhv2ef423b3tntgh2zjk5mxe1aogd65ztslwukgfyce
removed bafkreitgdnh6upcw5n16t2vm6epde5mez6yusvd4fpl4loll6p2koyca
removed bafkreitedb2xcgkyabxnrc17qnb3zuuqhc6qnaugqont4kdavppwr42lca
removed bafkreialjxowfywtst5w56ohx3rf3fx446uvacd2vbfj1nbgghjklqxu
removed bafkreif4shyo5wul74eep74vs4ggfqnhd4llwqr74glbrrdbdebpb3qxu
removed bafkreitegdcvj1leyu36xrh5g6fkeqp6036tub4uon7dbx1x7gwm2jg7fm
removed bafkreiddec2sl4mq44jln5nfjez3znqln7xlox5zxbvbn27onkwudj1rwm
removed bafkreifnbh3phnxxw1j3dxrrs44cfx3g7ovwxy3mffn674m45uuugwe1
removed bafkreid2rq12jluzsvs673wglz1bd46hcln44xrtw3kdaes36xh6gfge1
removed bafkreiflzd76pqz12zwo7dnq6shjlsbss7mft7ndyyoyntcqlvycpiyobm
removed bafkrethpxfhdnm3s4hpfvfwkn7hlh1xuazbzmtsg7r4jj37oflcav7k6bl
removed bafkrethyyocwghnlok3tdf2ebnmcuuxn47l5jt6vmcyhkl1apmzksoba
removed bafkreitbl7thhgeq26eetsnjyd7c2uq5cvjy3l56x2g7qus7xa3fvd1ba1
removed bafkreitjuxy5f63yuptq6m5iskyzupxwzi1b1heg5o6hsz6qbrkf5kba
removed bafkreigt7mq6ylobpepk2sgpjx4eq3hqudow2y3j3sttkzlc46j4utj6cgy
removed bafkrethvt6zbo3yha67mxjxc2kjlz6bg5fo6lg5lgyt2dvuxzlaotdmcgy
removed bafkreitapqac6x5ob47gztmo5g5qztozrmxge24sq4le3b43rdzgx3h3ay
removed bafkrethf7ybmktclx1k5ah663k6lfttjine3g3hs2nebk2pltwqnl6zay
removed bafkreitgz7r5az5yongb46o6njuo3g5z7mw3zdj3yufq1q5oetqobgmue
removed bafkreitgmom7xgevh1husznza3dvvbow42cxtvxqrweaqmhvegnuy1hxaqy
removed bafkreitaa7vm47dverShoqneocpt3tabze76tnkbnr1ctkl12hzsazw2qqy
removed bafkreidfkvbybmbchzzl1tr7tjwhqb66drqs63lqchsbfcnh5jk2vvrqqy
removed bafkreid4ztgrlfsckqv26psopnm236awmcgmgsuzmj3s3tpgwml14wece
removed bafkreitclt5zjkr2ap7om1545fuplfwhyewjp5ztq6vlinsnjugfnnrhhfucm
removed bafkreitbg6f26qzjsn76l5udlnqkjnex7w76ze4olt5gquathlqwgxeecm
removed bafkreitag7htaztonmaq26p36nlvkh7zx3ydn4hhoc7jaoug14hqf6ece
removed bafkreicuygpbkwnqr1sf4xy75zyaselfpsyx72j4hd5umjpdfdjjx5vm
removed bafkreitergqx74ojbc65veospgonh75cdcon2qdhagfgxal2m2de237oaye
removed bafkreith3r6vw7stycupsnpv4ptoyezaxhen6etrgvalnachqt2fhte2ppq
removed bafkreitlangko1pk4fz76mgzjlfkasugzlooyrpvx7zmtlatzdnjzfyxt5e
removed bafkreitbpapf2pdfjvq7tbmasti7tuvugvdpmuejp37kvwgwnwnc32tht5e
removed bafkreitavcax4lkhfaha3eadsx56wks3whfefnoa4s5nc775233tf5o7gjr4
removed bafkreithntgo5cdoqjva5v2b2fyyogywys2lqnkqbqznulqguvuxzdh4r4

```

Figure 76 Garbage collection to delete unpinned content.

- Figure 77 shows the integration of KR with the Desktop to utilise services such as knowledge upload, peer visibility, explore stored contents, etc.

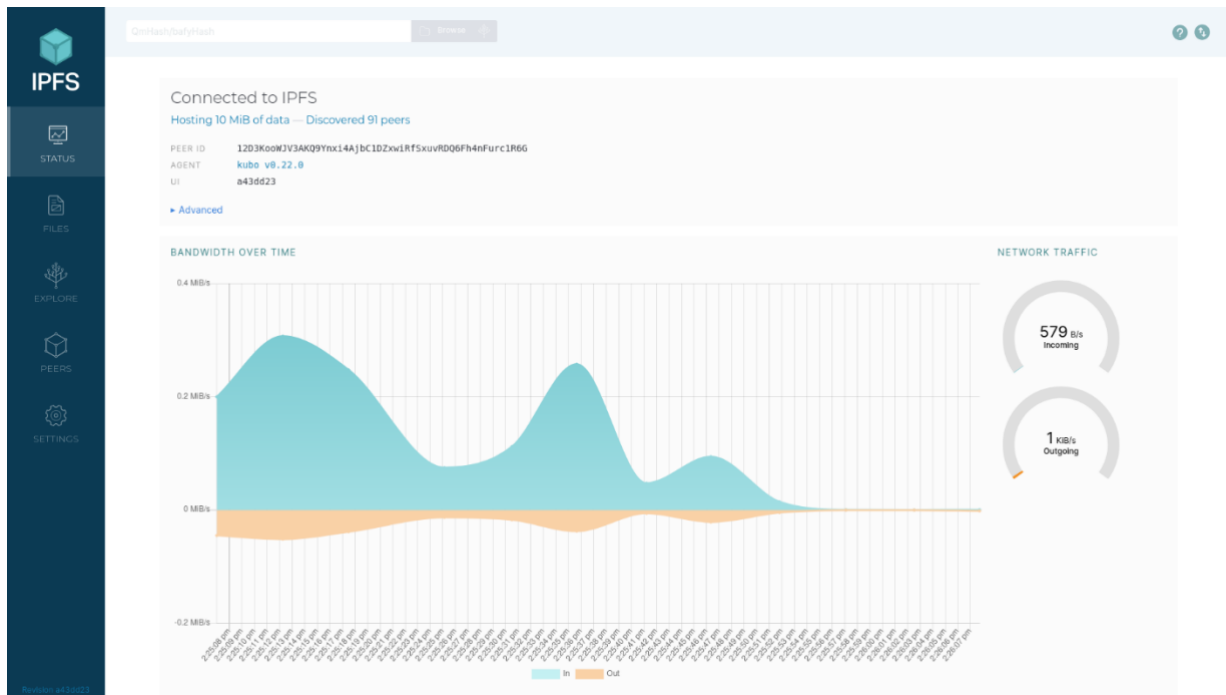


Figure 77 IPFS desktop integration to KR to avail IPFS services.

7.7. NEXT STEPS

The next steps that will be followed to accomplish the implementation goals are as follows;

- Integrate KR with ATR for immutable reference and efficient search of the CID based on the metadata.
- Measure the performance of KR.
- Implement the Automated Analysis functionalities to the stored knowledge content.
- Implement Knowledge Graph functionalities to enable seamless access to the content stored in the KR.

7.8. CONCLUSIONS

In Conclusion, the focus of Task 3.4 is to implement the Knowledge Repository utilising IPFS. The IPFS has been proven to be one of the most promising solutions when it comes to knowledge management and distribution. In addition, it comprises several security enhancements, such as ensuring a tamper-resistant repository, comprising cryptographic features, safeguarding against unauthorised access and knowledge manipulation and so on. Due to its innovative features of IPFS, the TRUSTEE KR will be empowered with innovative methods compared to traditional repositories. While the KR will play a pivotal role in reshaping how the knowledge-based contents are stored, shared and accessed within the TRUSTEE platform, it will also ensure faster knowledge retrieval times, reduced latency and improved resilience.

8. PART VI: AUTHENTICATION MANAGER (AM)

8.1. INTRODUCTION

The purpose of this section is to analyse the architecture of the Authentication Manager (AM). This module's objective is to develop, implement and utilise a Self-Sovereign Identity (SSI) based on blockchain technology to allow to authenticate TRUSTEE users and bridging our system with the eIDAS network for retrieving a user's identity information from a trusted source. In particular the SSI technology will be used for authenticating users while the eIDAS regulatory framework will be used to issue a user's identity by retrieve data and information through eIDAS, making it available to the user by issuing him/her verifiable credentials to be stored and managed on a user's wallet.

Additionally, the AM component will be extended with support to GAIA-X users (in collaboration with T4.4), who already have their own GAIA-X credentials that could be directly considered in TRUSTEE. For this purpose, wallet.id technology will be considered. This extension could be considered as the basis for other types of users. This section will start by providing background information on authentication mechanics, user authorisation, the SSI concept and briefly mention state-of-the-art mechanics. On the second part, the high-level architecture of the Authentication Manager (AM) will be explained and then move into the description of the detailed architecture.

8.1.1. NEED AND OBJECTIVES

Authentication plays a vital role in securing information systems, ensuring that only authorized entities are granted access to a system's resources. As technology continues to evolve, the field of authentication has witnessed significant advancements, shifting from traditional password-based approaches to more innovative and secure password-less authentication mechanics.

The concept of authentication emerged with the rise of digital systems and the need to establish trust in online interactions. Traditionally, authentication primarily relied on knowledge-based methods, such as passwords or personal identification numbers (PINs), which required users to remember a secret piece of information and present it upon request to prove that they know it. However, the proliferation of password-related attacks, such as brute-force attacks on weak passwords, highlighted the inherent vulnerabilities of such password-based approaches. Consequently, the need for more robust and user-friendly authentication mechanisms became apparent, leading to the exploration of alternative methods like biometrics using tokens and cryptographic keys.

The current state-of-the-art authentication revolves around upgrading the security of traditional password-based authentication systems by deploying multifactor or strong password-less authentication mechanics. Password-less authentication has gained considerable attention due to its potential to address a number of weaknesses associated with passwords. This approach typically employs a combination of biometrics, hardware tokens, or mobile devices to verify user identity. Biometric authentication techniques, such as fingerprint recognition, facial recognition, or iris scanning, offer increased convenience and security. Hardware tokens, such as security keys or smart cards, provide an additional layer of protection against phishing attacks.

8.1.2. INNOVATION

Self-Sovereign Identity (SSI) represents an emerging paradigm in authentication that aims to empower individuals with full control over their personal information and digital identities. SSI leverages blockchain technology to enable individuals to manage their identity attributes securely on their wallets, while selectively sharing them with trusted parties, without the need for intermediaries or central authorities. SSI eliminates the need for users to rely on a single, trusted third party to authenticate their identities, thereby reducing the risks associated with centralized identity management systems. By leveraging cryptographic protocols, SSI enables users to prove their identity without disclosing unnecessary personal information, enhancing privacy and security in the digital realm.

Combining the total control of one's identity through SSI and securing it using strong password-less authentication mechanics not only provides immediate benefits for the user but also paves the way for the future of authentication. This innovative approach has the potential to revolutionize the landscape of digital identity management, addressing the limitations of traditional authentication methods and setting new standards for trust and security in the digital realm.

8.2. BACKGROUND

8.2.1. SELF-SOVEREIGN IDENTITY (SSI)

In the concept of Self-Sovereign Identity (SSI), the management of the identity of the user and its personal data are given to the user itself. Users are able to acquire proofs of their attributes (e.g., proof of successful completion of University programme), issued to them by a trusted entity (e.g., by a university) and present them when asked for to 3rd parties (e.g., present proof to a recruitment agent during a job interview) without revealing anything more than the proof itself.

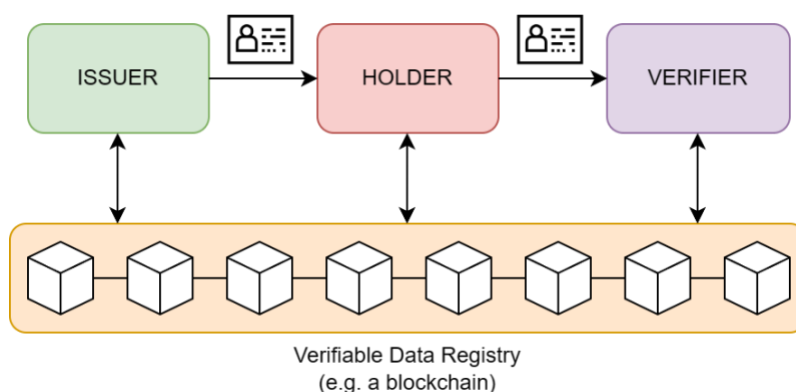


Figure 78: Overview of the 3 main entities of SSI.

8.2.2. SSI IN GAIA-X

GAIA-X, also known as the European data infrastructure project, is a visionary initiative that aims to establish a federated data infrastructure for Europe. The project's goal is to create a trusted, sovereign, and decentralized data infrastructure that enables secure data sharing and interoperability across various sectors and countries. At its core, GAIA-X seeks to provide a framework for building a European data ecosystem based on principles of transparency, interoperability, data sovereignty, and trustworthiness. By doing so, it aims to address the current challenges associated with data fragmentation, vendor lock-in, and lack of control over data. GAIA-X envisions the creation of a decentralized and secure data infrastructure that enables data sharing and collaboration while respecting privacy and security requirements. By fostering interoperability between various cloud providers, data services, and infrastructure providers, GAIA-X aims to simplify data access, enhance data portability, and facilitate the development of innovative services and applications.

SSI plays an important role in establishing trust between entities within GAIA-X ecosystem. By enabling verifiable credentials and cryptographic proofs, SSI offers a secure and tamper-proof way to authenticate and verify identities. This trust foundation is essential for enabling collaboration, interoperability, and cross-border data exchange within GAIA-X while maintaining data sovereignty and protection. In the context of GAIA -X, SSI can have numerous applications for identity and trust, which are listed below:

- Self-Issued OpenIDProvider (SIOP) for decentralised authentication.
- Attribute-based authorisation with credentials over role-based access control.
- Self-management of credentials through secure wallets for individuals and organisations.
- Software agent for selective disclosure based on verifiable credentials.

Although there are many SSI technologies that allow working with verifiable credentials, Walt.id has been considered to store and manage verifiable credentials corresponding to GAIA-X. GAIA-X users use specific verifiable credentials with specific attributes, which will be referred in the document as GAIA-X credentials. Two examples of GAIA-X credentials are:

- GAIA -X association membership.
- GAIA -X Working Group membership.

8.3. REQUIREMENTS

In this chapter, the compliance of the AM ABB of TRUSTEE is investigated with regards to the user-centred, technical, legal, and socio-ethical requirements that were reported in D2.1 (TRUSTEE consortium, 2023). The results of this investigation provide further preliminary insight regarding with the aforementioned requirements addressed during the development and by the implementation of the AM and are presented in the following sub-sections. It is worth noting that only the Mandatory, Desirable, and Optional requirements have been considered at this point. Furthermore, only the currently compliant or partially-compliant requirements have been considered. As the implementation of the AM progresses during the development of the TRUSTEE Platform and the maturity of the ABB increases, the investigation of the compliance will evolve and further updates and/or enhancements will be provided in upcoming versions of this deliverable, namely D3.2 and D3.3.

User-centred requirements were identified through the UX Research that was conducted under T2.3. These requirements were reported in D2.1 (TRUSTEE consortium, 2023), while they were also utilized in order to extract technical, functional and non-functional requirements, that were also outlined in D2.1. Table 22 below presents the interplay between user-centred requirements and technical requirements, that is the functional and non-functional requirements that address each user-centred requirement. The user-centred requirement ID is used to reference the respective requirement reported in D2.1 (TRUSTEE consortium, 2023) while the requirement's description is also provided alongside whether the functionality is currently fully or partially compliant with the user-centred requirement, i.e., indicating whether technical requirements have been identified in order to address the respective user-centred requirement. The ID of the technical requirement, as reported in D2.1 (TRUSTEE consortium, 2023), that is addressing each user-centred requirement is also provided in the last column of the table. Additionally, it is possible than more than one technical requirement addresses a single user-centred requirement.

Table 22: AM – User-centred requirements

Req id (WP2)	Req. Description	Compliance	How is addressed
USR-003	Users must provide access to their data in a secure and privacy aware way without sending their raw data to others	Y	Req-AM-FUNC-1
			Req-AM-FUNC-5
			Req-AM-NFUNC-1
			Req-AM-NFUNC-2
USR-005	The data provided must be used with respect to privacy regulations, legal	Partially	Req-AM-FUNC-1

	frameworks, and ethical considerations and without the risk of data loss		Req-AM-FUNC-2 Req-AM-FUNC-3 Req-AM-FUNC-4
USR-008	Only authenticated users must have access to the data of the providers	Partially	Req-AM-FUNC-1 Req-AM-FUNC-2 Req-AM-FUNC-3 (also to be supported by developments in T4.6)
USR-011	TRUSTEE must enable registration and creation of a new account for each user	Partially	Req-AM-FUNC-2 Req-AM-FUNC-3
USR-019	TRUSTEE should confirm the authenticity of the data in an easy way	Y	Req-AM-FUNC-2 Req-AM-FUNC-5 Req-AM-NFUNC-1 Req-AM-NFUNC-2
USR-039	TRUSTEE should be secure and free of cybersecurity issues	Partially	Req-AM-FUNC-1 Req-AM-FUNC-2 Req-AM-FUNC-3 Req-AM-FUNC-4
USR-044	TRUSTEE should allow data that are valid and provided by authenticated users	Y	Req-AM-FUNC-2 Req-AM-FUNC-5
USR-054	TRUSTEE should have easy access and permissions	Y	Req-AM-FUNC-2 Req-AM-NFUNC-1 Req-AM-NFUNC-2
USR-057	TRUSTEE should be an efficient and highly secure platform	Y	Req-AM-FUNC-1 Req-AM-FUNC-2 Req-AM-NFUNC-1 Req-AM-NFUNC-2
USR-100	The user could be able to decide the level of access they are willing to provide for their data	Y	Req-AM-FUNC-1

Req-AM-FUNC-2

Table 23 presents the compliance of the Authentication Manager with the Technical functional and non-functional requirements based on a preliminary investigation of the architecture and components current state. These technical requirements can be found on D2.1 using the requirement ID listed on the table. Further updates and/or enhancements regarding the compliance with the technical requirements will be provided in future versions of this document (namely D3.2, D3.3) as the implementation of the Authentication Manager progresses.

Table 23: AM – Functional and non-functional technical requirements

Req id (WP2)	Req. Description	Compliance	How is addressed
Req-AM-FUNC-1	TRUSTEE users/Personas have control of their identity and the data they provide to the platform.	Y	TRUSTEE Authentication Manager will allow users to select what information to share with the TRUSTEE platform.
Req-AM-FUNC-2	To access the various systems users/Personas will have to be authenticated.	Y	TRUSTEE Authentication Manager will allow ABBs to authenticate users through OIDC.
Req-AM-FUNC-3	The system will have to allow the registration of new users/Personas.	Partially	TRUSTEE Authentication Manager will allow the registration/onboarding of new users. (This is partially implemented and will be fully implemented when the eIDAS related functionality is mature enough.)
Req-AM-FUNC-4	TRUSTEE should provide strong user authentication mechanics.	Y	TRUSTEE Authentication Manager will use strong and user-friendly authentication mechanics.
Req-AM-FUNC-5	TRUSTEE should ensure the information provided by Data Providers and Consumers is trusted.	Y	TRUSTEE Authentication Manager will only use information sourced from parties or information signed by trusted parties.
Req-AM-NFUNC-1	TRUSTEE should provide efficient user authentication mechanics.	Y	TRUSTEE Authentication Manager will use strong and user-friendly authentication mechanics.
Req-AM-NFUNC-2	TRUSTEE should feature an easy-to-use authentication.	Y	TRUSTEE Authentication Manager will use strong and user-friendly authentication mechanics.

A preliminary analysis of the compliance with the legal and socio-ethical requirements defined in the D2.1 (TRUSTEE consortium, 2023) is shown in ANNEX 2 and ANNEX 3.

8.4. ARCHITECTURE

8.4.1. HIGH LEVEL

The Authentication Manager (AM) is one of the most essential components of TRUSTEE implementing the authentication and authorisation features needed by various TRUSTEE properties. Using the AM other ABBs will

be able to allow access or deny access to users depending on their identity attributes they are able to prove, providing in this way an effective and required access control on the TRUSTEE ecosystem. Multiple TRUSTEE ABBs (as indicated by Figure 79) will be interacting with the Authentication Manager to recover the identity of the user through its external APIs. Furthermore, users will also be interacting with the Authentication Manager to be authenticated as TRUSTEE Users using their SSI wallet (e.g. a compatible mobile phone using a compatible wallet application). The high-level architecture of the AM, which is based on the latest TRUSTEE system architecture, is presented on Figure 79.

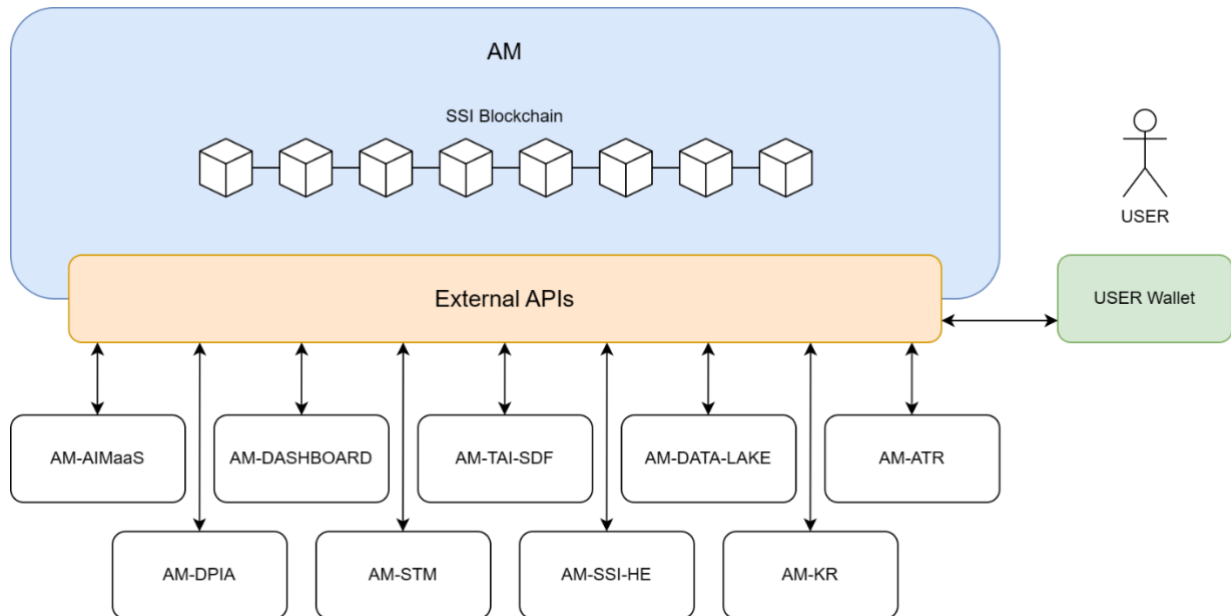


Figure 79: High-level architecture of the Authentication Manager interaction with other TRUSTEE related components.

Authenticator Manager's external APIs uses a number of protocols depending on the service to be provided with the external entity. As presented on Figure 80, other TRUSTEE ABBs (such as the AIMaaS or the Dashboard) will have to use the OpenID Connect API (based on OAuth 2.0) exposed by AM to authenticate users and recover the user's identity. As part of the authentication process, the user's client (e.g. a browser on the user's laptop) will communicate through HTTP(s) with the AM accessing it as a web application. The user will be then asked to prove his identity link its SSI wallet with the AM through DIDComm (e.g. by scanning an SSI invitation QR code) so that it can exchange information such as storing credentials, issued by the AM, inside the wallet (e.g. by inspecting and accepting the issued credentials and instructing their wallet the store them on their device) or providing proof of its attributes by sending to the AM the requested credentials proofs (e.g. by selecting on the wallet the compatible attribute they want to share with the AM, such as the name of the organisation the user is working for).

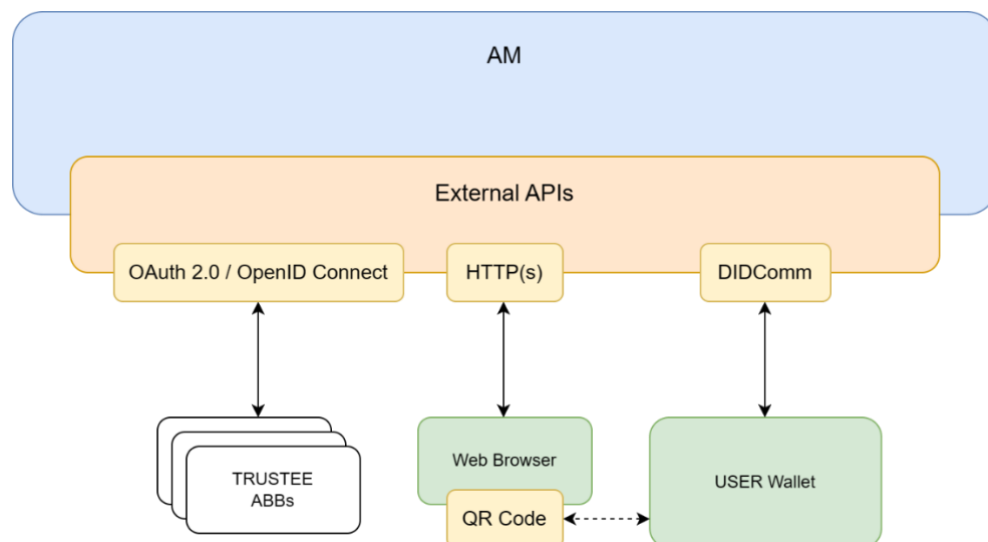


Figure 80: The external APIs and their protocols as exposed by the Authentication Manager.

8.4.2. LOW LEVEL

To achieve the needed functionalities, several components are needed inside the Authentication Manager (AM). In this section we will describe the purpose of each component and provide implementation related information. Figure 81 showcase a diagram of the Authentication Manager's detailed architecture.

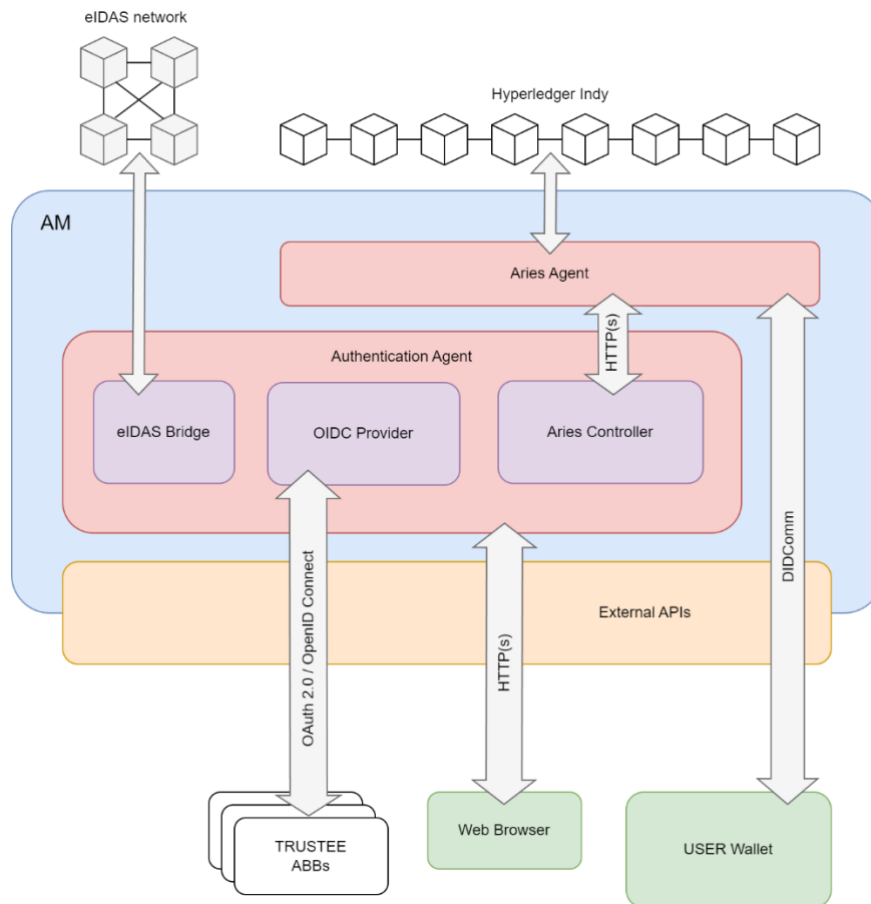


Figure 81: Detailed architecture of the Authentication Manager showing all the internal components.

Authentication Agent. One of the most important components of the AM is the Authentication Agent. By designed we chose to build the Authentication Agent as a web site to facilitate easy access for the user using a web browser, cross-platform compatibility by relying on widely adopted web standards, interoperability with other components in a technology independent way by utilising the popular OIDC. This web server is hosting the main control logic of the Authentication Manager while also allows the creation of user familiar interfaces (in the form of a web application) for interacting with them. The implementation of the Authentication Agent will be based on Python to minimise possible limitations in the ability to develop new functionalities and on the popular Flask⁷ web framework to maintain flexibility and ease of use. The Authentication Agent wraps inside the OIDC Provider, the Aries Controller and the eIDAS bridge subcomponents that will be analysed further in the following sections.

OIDC Provider. In order to bring the SSI concept to the web environment and allow its seaming-less use by external ABBs without the need to implement complex SSI related procedures (e.g. require them to implement an SSI agent) or interact with the blockchain, the popular OpenID Connect was selected as a means to start the authentication process and transferee the user's identity to other ABBs when needed. The OIDC Provider subcomponent, lives inside the Authentication Agent implementing and exposing the appropriate OAuth 2.0 and OIDC endpoints that authorised external applications will be able to use during authentication. For our

⁷ <https://flask.palletsprojects.com/>

implementation we mainly focused on implementing the Authorization Code Flow of OAuth2.0/OIDC⁸. This was achieved using the oauthlib⁹ library.

Aries Controller. The Aries Controller is the subcomponent responsible for controlling the Aries Agent and coordinating the overall SSI related actions, deployed inside the Authentication Agent. The Aries Controller communicates with the deployed Aries Agent over HTTP using its administrator API¹⁰. Implemented in Python as part of the Authentication Agent, the controller initialises the connections between the user's wallet agent and instructs the local Aries Agent on what actions to perform (e.g., accept connections, request credential proofs or issue credentials to the user) following the appropriate protocols (Aries RFC0160¹¹ for establishing a connection with the client, Aries RFC0037¹² for requesting and getting a presentation proof and Aries RFC0036¹³ for issuing credentials).

eIDAS Bridge. The eIDAS Bridge is the subcomponent responsible for retrieving the user's information from the eIDAS network and passing those information to the Aries Controller building a virtual bridge from the eIDAS ecosystem to the SSI ecosystem used by TRUSTEE. As eIDAS is a trusted source of information, Aries Controller will be able to issue to the user credentials based on the information retrieved by eIDAS Bridge.

Aries Agent. The Aries Agent component includes an instance of Hyperledger Aries Cloud Agent Python (ACA-Py)¹⁴ and is used as an SSI agent (acting as an issuer and a verifier) for interacting with the blockchain (agent-to-blockchain communication) as well as for interacting with other agents (agent-to-agent communication) such as the wallets of TRUSTEE users that want to be authenticated. The Aries Agent is being controlled by the Aries Controller instructing it on how to generate and handle connections, what credentials to request from the clients (TRUSTEE users) or what credentials to issue to them.

Hyperledger Indy. The Hyperledger Indy acts as the SSI blockchain. For the development this will be implemented using a series of Hyperledger Indy nodes. The nodes will be deployed with the help of the VON Network project¹⁵. The Aries Agent is interacting with this component to store/register or retrieve/validate SSI related information. Apart from AM's Aries Agent component, other SSI agents (e.g. the TRUSTEE user's wallet) may also communicate and interact with this blockchain for the same reasons. The ledger technology to be used is not affecting the architecture of the Authentication Manager, thus it may change depending on the project needs.

eIDAS Network. The authentication manager will have to talk with the eIDAS network (the eIDAS ecosystem) through the eIDAS bridge. This network features a number of eIDAS nodes that could be used to retrieve the user's identity and authenticated him using his eIDAS compatible ID.

8.4.2.1. EXTENSION FOR GAIA-X USERS

GAIA-X credentials are W3C-formatted, while Hyperledger Aries works mainly with Hyperledger AnonCreds, although it is currently in the process of supporting also W3C credentials. The purpose is to later use those Verifiable Credentials to access data or service that require access control. The access will be controlled using attributes from

⁸ https://openid.net/specs/openid-connect-basic-1_0.html

⁹ <https://github.com/oauthlib/oauthlib/>

¹⁰ <https://github.com/hyperledger/aries-cloudagent-python/blob/main/AdminAPI.md>

¹¹ <https://github.com/hyperledger/aries-rfcs/blob/main/features/0160-connection-protocol/README.md>

¹² <https://github.com/hyperledger/aries-rfcs/tree/main/features/0037-present-proof>

¹³ <https://github.com/hyperledger/aries-rfcs/tree/main/features/0036-issue-credential>

¹⁴ <https://github.com/hyperledger/aries-cloudagent-python>

¹⁵ <https://github.com/bcgov/von-network>

the Verifiable Credentials. GAIA-X has included support for walt.id so it is expected that this technology fits well with GAIA-X ecosystem. Furthermore, **walt.id** adds support for OpenID Connect (OIDC), as Keycloak, for example, does, so both technologies can be connected. Moreover, walt.id can work with many ledgers. Therefore, walt.id seems to be a more suitable alternative than Hyperledger Aries for working with GAIA-X credentials within TRUSTEE.

Walt.id provides both enterprise and consumer Wallets for SSI that adhere to the latest global and EU SSI standards. These wallets are licensed under Apache 2, a permissive open-source license. Walt.id has an IDP kit that allows you to initiate an identity provider that adheres to the OIDC standard. It utilizes the OIDC-SIOPv2¹⁶ (Self-Issued OpenID Provider v2) protocol and/or NFT blockchain APIs to obtain identity data or NFT metadata through a compatible wallet. Identity data stored in a Web3, or SSI wallet can be used as OIDC user info and/or mapped to standard OIDC claims.

Figure 82 represents the modular architecture of the authentication system designed for the authentication of different stakeholders that could use TRUSTEE platform; TRUSTEE users and GAIA-X users. The authentication system is designed to allow the identification of GAIA-X users, by presenting claims via W3C Verifiable Credentials, through a federated OpenID Connect (OIDC1) sign-in flow.

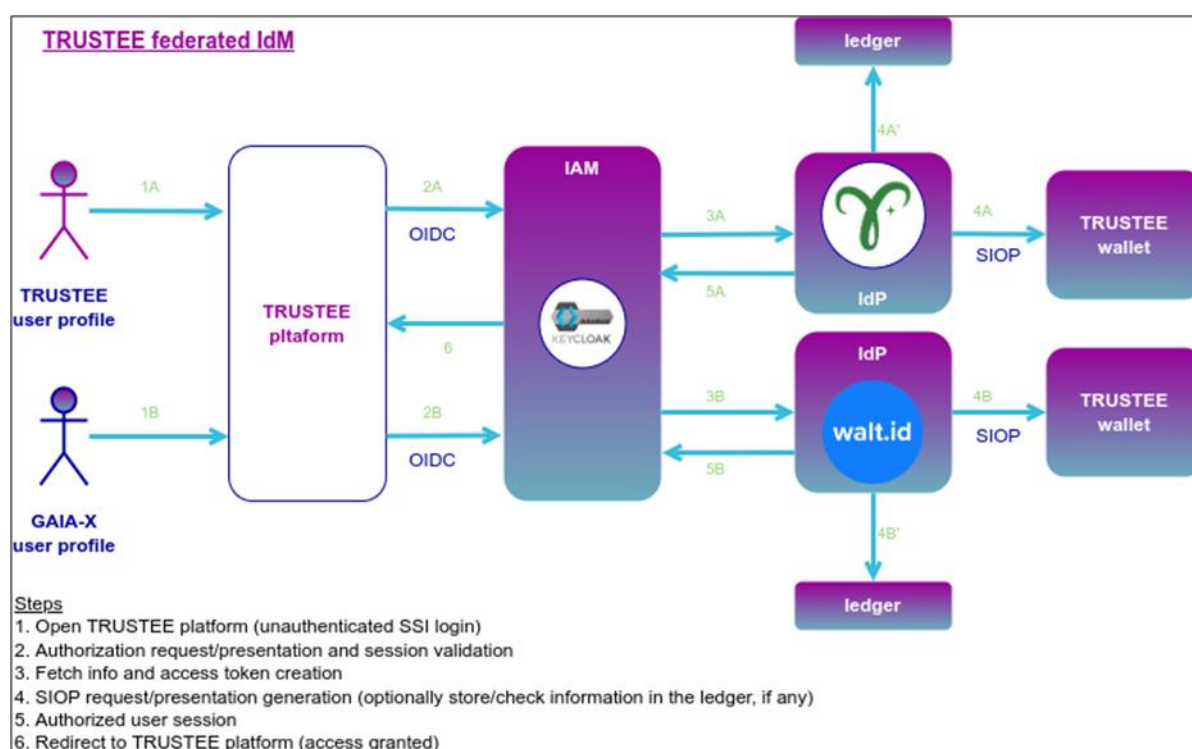


Figure 82: Authentication Manager extension for GAIA-X users context.

The main components are:

Keycloak Identity Access Management (IAM) Broker. While it may constitute an Identity Provider by itself, in TRUSTEE identity provisioning is federated. Keycloak can federate external user databases via Kerberos or LDAP,

¹⁶ https://openid.net/specs/openid-connect-self-issued-v2-1_0.html

or act as a broker with third party IdPs via OIDC or SAMLv2. In general, OIDC federation applies in our use cases. Keycloak can thus be considered our IAM broker.

SSI IdPs. The main objective of our architecture is to facilitate the authentication of stakeholders through the verifiable presentations kept by themselves in accordance with different SSI stacks. Thus, the SSI application layer can be said to represent the IdP in our system. Currently, in TRUSTEE Hyperledger Aries and walt.id have been considered. Both are application-layer SSI toolkits that facilitate the operation and interaction of the different actors involved in SSI authentication processes. Both can interoperate with different SSI ledger technologies. Two independent modules will be built, one for each of these two toolkits.

The SIOPv2 protocol enables OIDC compliant operation to these toolkits. In this way, the identity can be federated from the Keycloak IAM Broker to the SSI IdPs as for any other OIDC compliant IdP. Through SIOPv2, the result of the SSI credential proof verification process is mapped to JWT claims on the JWT token that authenticates the stakeholder.

In relation to the SSI IdP module walt.id, walt.id provides the IDP Kit¹⁷ for the supply of the SIOPv2 service. This kit is based on the WalletKit library (which in turn uses the SSKit library¹⁸). An instance of WalletKit¹⁹ is also required as a service for user wallet management and proof generation. Its SIOPv2 OIDC code flow is extensively described in walt.id documentation²⁰.

Ledgers. Blockchain DLTs (Distributed Ledger Technologies) endorse the DIDs of issuers and verifiers. They also keep track of the schemes that define verifiable credentials. This module encompasses both the distributed ledger itself and the framework that implements the SSI layer 1 protocol. Involved DLT technologies are Hyperledger Indy²¹ and IOTA²².

TRUSTEE Wallet. A secure web crypto wallet oversees the following services:

- Safeguarding of stakeholders' verifiable credentials. Access to verifiable credentials shall be granted only and exclusively to the sovereign credential holder.
- Acquisition of verifiable credentials from issuers.
- Submission of verifiable presentations to verifiers.

The development of the wallets will make use of APIs from the same application-layer SSI frameworks involved in SSI IdPs (e.g.: walt.id Wallet Kit).

The main benefits enabled by this approach are listed below:

- OIDC is a widely adopted standard authentication protocol. This makes it particularly suited as an entry point to the authentication system.
- The identity provider federation enables the integration of different identity verification methods: sign-in via user/password, 2-factor authentication, SSI verifiable presentations, etc. Not only different mechanisms, but also different IdPs, ledgers, etc. This opens the door to a heterogeneous and decentralized

¹⁷ <https://walt.id/idp-kit>

¹⁸ <https://walt.id/ssi-kit>

¹⁹ <https://walt.id/wallet-kit>

²⁰ <https://docs.walt.id/v/idpkit/concepts/identity-via-ssi>

²¹ <https://www.hyperledger.org/use/hyperledger-indy>

²² <https://www.iota.org/solutions/auditing-and-ssi-tools>

ecosystem where different authentication mechanisms and sources of truth coexist: legacy, present and future.

- Legacy authentication mechanisms may be integrated.
- SSI gives individuals full control over their credentials, securely held in their crypto-wallets.

The detailed architecture for GAIA-X users' authentication is shown in Figure 83. GAIA-X users manage their own *walt.id* wallet. At the same time, they can authenticate through a federated OpenID Connect (OIDC1) sign-in flow, as explained before. The *walt.id* wallet provides the identities to the Keycloak for this authentication. Optionally, the ledger is checked for getting public information about the issuer of the credential.

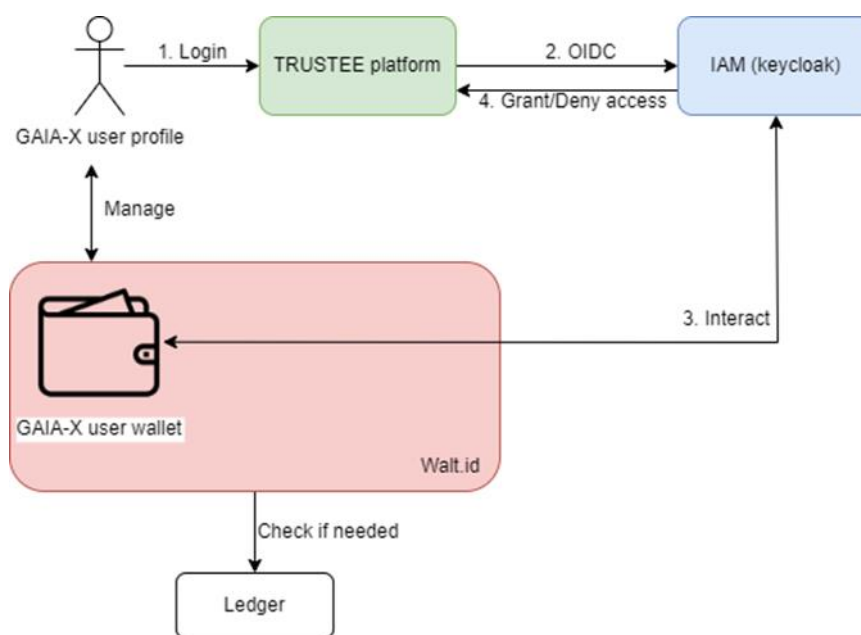


Figure 83: Authentication Manager extension for GAIA-X users architecture.

8.5. IMPEMENATATION

8.5.1. FUNCTIONAL DESCRIPTION

In this section we will look into functions and communication interfaces exposed by the Authentication Manager to communicate with other ABBs and 3rd parties. Table 24, presents one of the most important interface that allows multiple ABBs – namely the TRUSTEE Dashboard, the Security and Trust Manager, the Data Protection Impact Assessment, the AI Model as a Service, the SSI Homomorphic Capable Framework, the Data Lake and the Trustworthiness AI Support Framework – to authenticate users and enforce access control.

Table 24: External interface between the Authentication Manager and multiple other ABBs to be used for user authentication.

Interface	Content	Comment

Interface ID	IExt:AM:ABBAuthAgent	
Interface to	TRUSTEE Dashboard - DA Security and Trust Manager -STM Data Protection Impact Assessment - DPIA AI Models as a Service – AIMaaS SSI Homomorphic Capable Framework (SSI-HE) Data Lake – DL Trustworthiness AI Support Design Framework - TAI-SDF	<i>Many of the TRUSTEE ABBs will need to use this interface to authenticate users.</i>
Interface overview	The interface is associated to an Authentication Agent that as a software will be part of the ABBs which will require to verify that a user is already authenticated or requires to be authenticated before accessing the service, or to authenticate a user.	
Messaging pattern	Request/response	
Data exchanged via this interface	The needed attributes from the user's identity.	
Communication protocol	HTTP(S) / OIDC	
Message format	JSON	
Security and integrity		

The following Table 25 presents the external interface between the Authentication Manager (AM) and the Accountable Transactions Recorder (ATR) in order to send information that needs to recorder for accountability or security reasons.

Table 25: External interface between the Authentication Manager and the Accountable Transactions Recorder.

Interface	Content	Comment

Interface ID	IExt:AM:ATR	
Interface to	Accountable Transactions Recorder – ATR	
Interface overview	This interface will be used for logging AM transactions to the ATR ABB	
Messaging pattern	<i>Request/response, publish-subscribe, notification</i>	
Data exchanged via this interface	<i>Transaction data</i>	
Communication protocol	<i>e.g., gRPC, HTTP(S), REST API, FABRIC API</i>	
Message format	<i>JSON, XML, CSV, protobuf, CBOR, BSON.</i>	
Security and integrity		

Another important external interface is the one needed for exchanging information between SSI Agents. This interface is important for connecting a user's wallet with the Authentication Manager in order to retrieve its issued credentials or provide proof of its identity. This interface is listed on Table 26.

Table 26: External interface between the Authentication Manager and external SSI Agents (e.g. a user's wallet).

Interface	Content	Comment
Interface ID	IExt:AM:Agent2Agent	
Interface to	TRUSTEE AM Agent External SSI Agent	
Interface overview	Communication between 2 TRUSTEE SSI Agents. Depending on the agent and the operation to be performed this interface could facilitate various activities (e.g. messaging, presentation of attributes, issue of verifiable credentials)	
Messaging pattern	Request/response	

Data exchanged via this interface	DID data	
Communication protocol	DIDComm/HTTP(S)	Authentication Agents use the DIDComm protocol to communicate between each other. DIDComm can be transmitted over any channel (e.g. HTTPS)
Message format	DIDComm	
Security and integrity		

Lastly but not least, an external interface linking the Authentication Manager with the external eIDAS network for retrieving a user's information from a trusted source and onboarding him/her on the TRUSTEE ecosystem. The information of this external interface with eIDAS is presented on Table 27.

Table 27: External interface between the Authentication Manager and the eIDAS network.

Interface	Content	Comment
Interface ID	IExt:AM:SSIEIDAS	
Interface to	SSI eIDAS Trusted Issuer	
Interface overview	The service responsible for linking SSI identities with the eIDAS will have to communicate with the eIDAS network. For this, a connection will have to be made towards the appropriate eIDAS node.	
Messaging pattern	Request/response	
Data exchanged via this interface	The needed attributes from the user's identity.	
Communication protocol	SAML	
Message format	eIDAS SAML Message Format	

Security and integrity		

8.5.1. OPERATIONAL DESCRIPTION

8.5.1.1. REGISTERING OIDC APPLICATIONS

During the setup/initialisation phase of the Authentication Manager, the applications that will use the exposed OIDC endpoints (e.g., ABBs that will would like to be able to authenticate users) will have to be registered on the OIDC Provider of the Authentication Manager. To register such an application client, the following information has to be provided to the Authentication Manager:

- Client Name (e.g., “TRUSTEE Dashboard”), this information will be visible to the users informing them which application is requesting them to be authenticated.
- Client ID (e.g., “1122334455667788”), this will be used to identify which application is requesting authentication.
- Client Secret (e.g., “11223344556677889900aabbccddeeff11223344556677889900aabbccddeeff”), this is a shared secret used to authenticate requests coming from a particular application (based on their Client ID)
- Client URI (e.g., “https://dashboard.trustee-dev.eu”), used to know the URI of the application.
- Client Redirect URI (e.g., “https://dashboard.trustee-dev.eu/oidc”), the URI to redirect the user to after the authentication completes (OIDC related parameters will pass on the request).

In the latest version of the Authentication Manager, the registration of the OIDC clients are considered to be known and agreed between the AM and other ABBs, and they are added directly in the OIDC Provider’s database during initialisation phase. From this point on, we will assume that the ABBs contacting the AM know their Client ID, Client Secret and the rest of the information is already known by the AM’s OIDC Provider component.

8.5.1.2. AUTHENTICATING A USER

Stating with the steps needed to authenticate a user A, TRUSTEE ABB that wants to authenticate a user will have the implement the OpenID Connect (OIDC) Authorisation code flow (which is based on OAuth2). As described by the OpenID Connect Basic Client Implementer's Guide²³, the steps to perform an Authorisation code flow are simple:

1. The client (in our case a TRUSTEE ABB looking to authenticate a user) prepares an Authentication Request containing the desired request parameters (namely “client_id”, “scope”, “response_type”, “nonce”). During this step the client will have to:
 - a. Know the OAuth2 authorise endpoint of the AM, which will look like “https://authentication-manager.trustee-dev.eu/oauth/authorize”

²³ https://openid.net/specs/openid-connect-basic-1_0.html

- b. Know its Client ID (generated during its registration as explained on the previous section)
 - c. Generate a new “nonce” bind to this request
 - d. Set the “response_type” to “code” to indicate an Authorisation code flow request.
 - e. Set the “scope” to “openid+profile”. This will later change to allow requests for specific verifiable credentials.
2. Client sends the request to the Authorization Server.
 - a. This can be done by redirecting the user’s agent to the Authorisation endpoint of the AM supplying the parameters specified on the previous step. For example, redirecting the user to: `“http://authentication-manager.trustee-dev.eu/oauth/authorize?client_id=1122334455667788&scope=openid+profile&response_type=code&nonce=abcdef”`
3. Authorization Server authenticates the End-User.
 - a. This step will be handled by the Authentication Manager’s SSI related components, namely the Aries Controller, Aries agent and the deployed blockchain.
4. Authorization Server obtains End-User Consent/Authorization.
 - a. This step will be handled by the Authentication Manager’s SSI related components, namely the Aries Controller, Aries agent and the deployed blockchain.
5. Authorization Server sends the End-User back to the Client with code.
 - a. The OIDC Provider component will redirect the user to the Client Redirect URI (that was set during the registration of the ABB on the AM as explained on the previous section) and pass the “code” parameter. For example, the AM will redirect the user to the: `“https://dashboard.trustee-dev.eu/oidc?code=1234567890abcdef1234567890abcdef”`
6. Client sends the code to the Token Endpoint to receive an Access Token and ID Token in the response.
 - a. The ABB will be able to recover the “code” parameter passed and use it to create a request towards the OIDC Provider’s token endpoint. To do so, the ABB will have to do a post request, using both its Client ID and its Client Secret (in order to be authenticated) towards the “/oauth/token” endpoint of the Authentication Manager posting the parameters “grant_type=authorization_code” and the code value retrieved on the previous step. Here is an example such request using the curl command in which the server will respond by providing an access token:

```
curl -u "1122334455667788:11223344556677889900aabbccddeeff11223344556677889900aabbccddeeff" -XPOST https://authentication-manager.trustee-dev.eu/oauth/token -F grant_type=authorization_code -F code=1234567890abcdef1234567890abcdef
```
7. Client validates the tokens and retrieves the End-User's Subject Identifier.
 - a. The ABB will then be able to retrieve the user’s information token (data provided by the SSI authentication). Assuming the AM server responding on the previous step providing the access token “abcdefabcdefabcdefabcdef”, here is an example request using curl:

```
curl -H "Authorization: Bearer abcdefabcdefabcdefabcdef" https://authentication-  
manager.trustee-dev.eu/oauth/userinfo
```

This whole process is presented in detail in Figure 84 including the here excluded steps to authenticate the user using its SSI verifiable credentials.

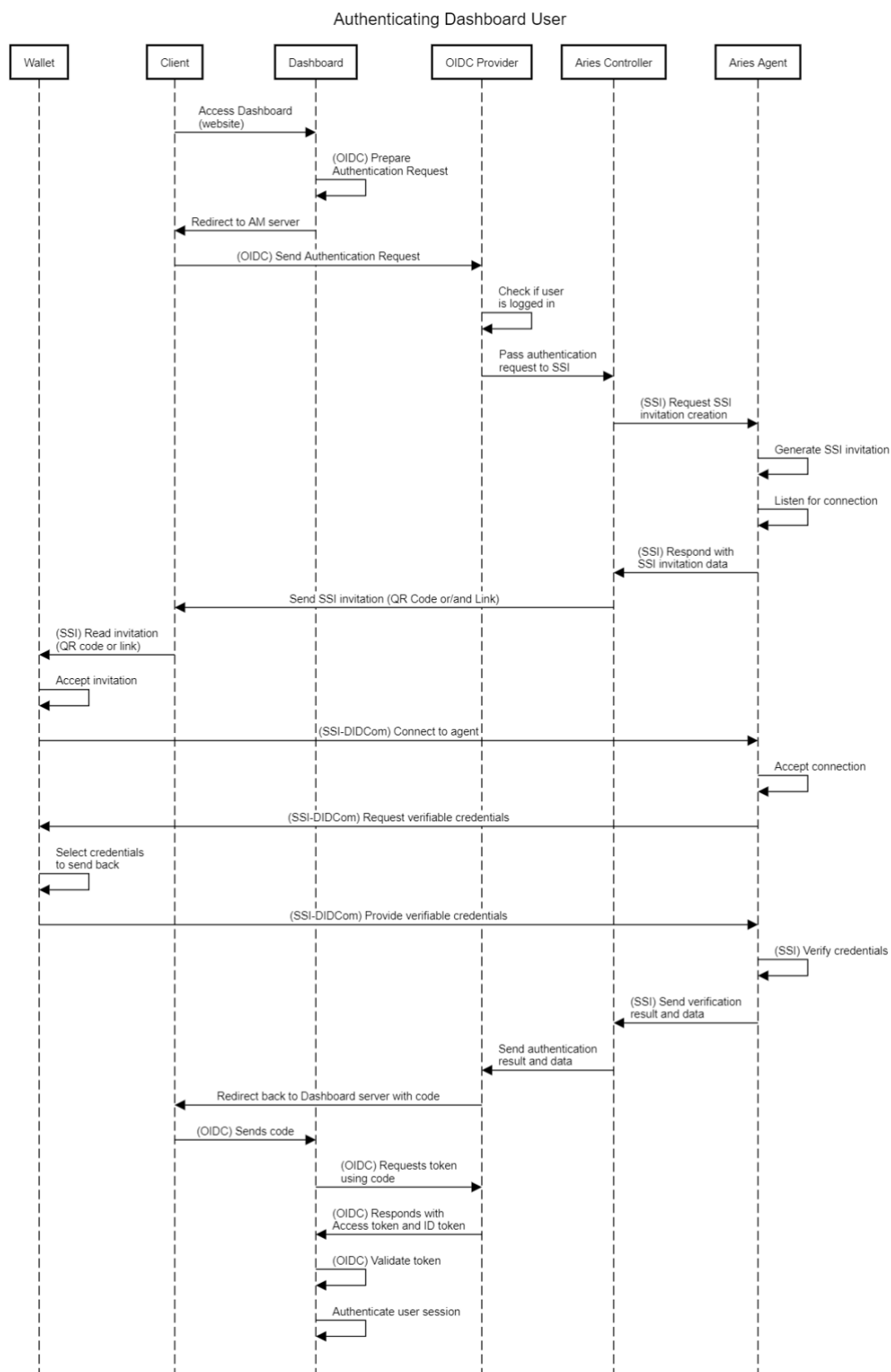


Figure 84: Example showing the step-by-step process of authenticating a user, showing the use of both OIDC and SSI authentication.

8.5.1.3. REGISTERING A USER

To be able to authenticate, the user will have to possess SSI credentials so that he can provide the requested attributes or prove the corresponding proof requests. A user may acquire such credentials from a trusted 3rd party (such as a university). In our case, as a first step to progress with the development of the Authenticator Manager and to facilitate the components to be implemented in the future, we implemented the credentials issue functionality inside our authentication manager. Thus, in the current state of the implementation, the identity information of the user will be taken directly from the user and the process of issuing SSI credentials will be launched by the Aries Controller. Assuming that the user does not hold any credentials inside his/her wallet that can be used to be authenticated on TRUSTEE, the following process should be followed so that he/she can receive new TRUSTEE related credentials to get access to the system:

1. The user will follow the authentication steps as presented in the previous section up to the point where the Authorization Server will have to authenticate the End-User (step 3).
2. Then, the user will have to select to register on the system by requesting to be issued TRUSTEE user's related credentials. To do so the user will start an authentication through eIDAS (selecting the appropriate option from the interface).
3. The user will then be authenticated through the eIDAS ecosystem (e.g. using the user's eIDAS compatible identity card)
4. The user will return on the Authentication Manager's website along with his identity information.
5. The Aries Controller of the AM will invite the user to connect his wallet with the Aries Agent (e.g. by scanning an invitation QR code).
6. Upon accepting the invitation for his/her wallet, the Aries Controller will instruct Aries Agent to send a credential issue request which the user will have to accept.
7. Following that the Aries Agent will process with issuing the credentials using the appropriate pieces of the identity information retrieved through the eIDAS network.
8. The user will then be able to store the issued credentials and use them to authenticate inside the TRUSTEE ecosystem.

This process is presented in detail in the following Figure 85.

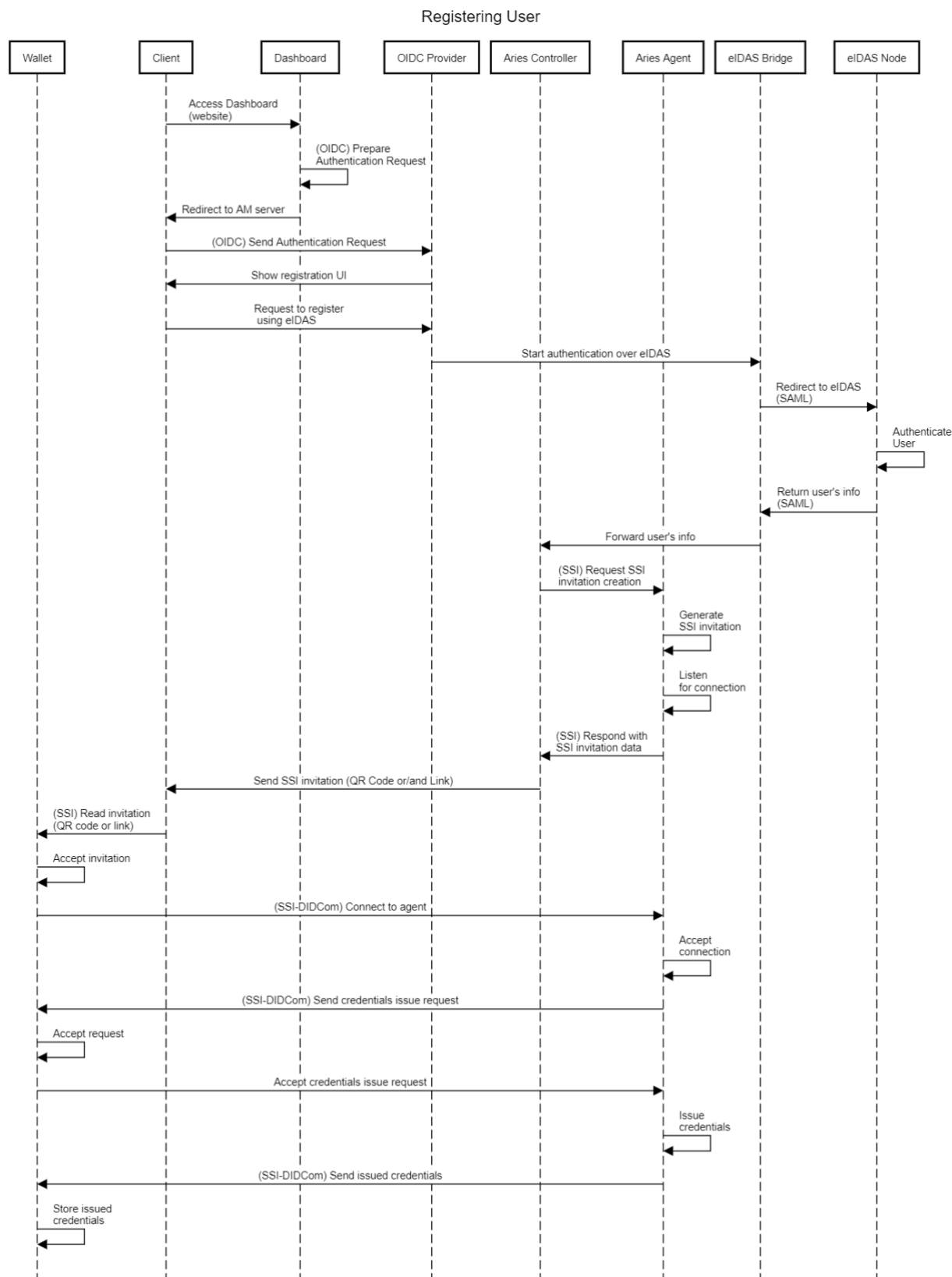


Figure 85: Example showing the step-by-step process of onboarding a user, showing the use of both SSI and eIDAS.

8.5.2. TECHNICAL DESCRIPTION

In this section we will explain the functionalities of the components and the subcomponents of the Authentication Manager. To better understand them we will describe them by analysing them through use cases.

8.5.2.1. REGISTERING OIDC APPLICATIONS

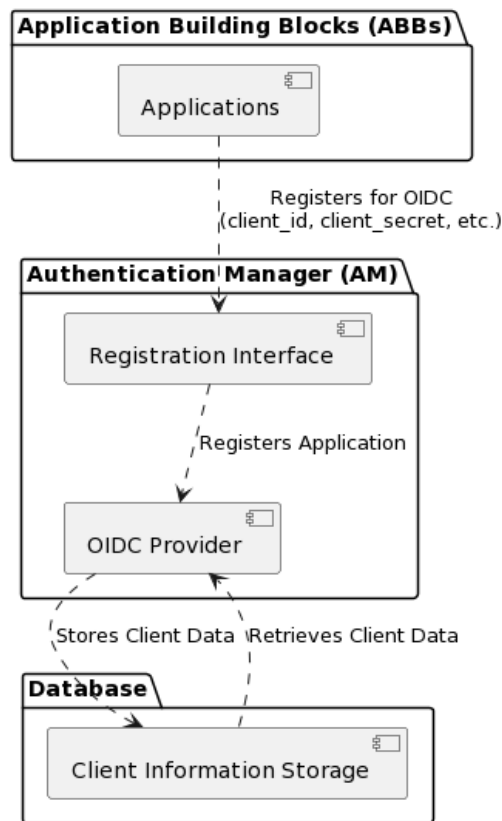


Figure 86 Block diagram illustrating the functionalities involved in registering OIDC applications.

The block diagram of Figure 86 visually represents the process and relationships involved in registering OIDC (OpenID Connect) applications, as described in your provided text. Here's a detailed explanation of each component and the connections shown in the diagram:

- **Authentication Manager (AM):** This is the central block in the diagram, representing the overarching system that manages the authentication processes. It oversees the entire flow of registering OIDC applications and ensures the correct implementation of the OpenID Connect protocol.
- **OIDC Provider:** Located within the Authentication Manager, this component is responsible for interfacing with the applications that require user authentication. It acts as the mediator between the applications (ABBs) and the database where client information is stored.
- **Applications (ABBs):** These blocks represent various applications that need to use the OIDC endpoints for user authentication. The diagram shows these applications connecting to the OIDC Provider, indicating the registration process where these applications provide their details (like Client Name, Client ID, Client Secret, etc.) to the OIDC Provider.
- **Database:** This block symbolizes the storage system where the OIDC clients' information is kept. It holds essential details such as the Client ID and Client Secret, which are necessary for the authentication process.

There is a direct connection from the OIDC Provider to both the Applications (ABBs) and the Database, signifying that the OIDC Provider receives information from the applications and stores or retrieves relevant data from the database.

8.5.2.1. AUTHENTICATING A USER

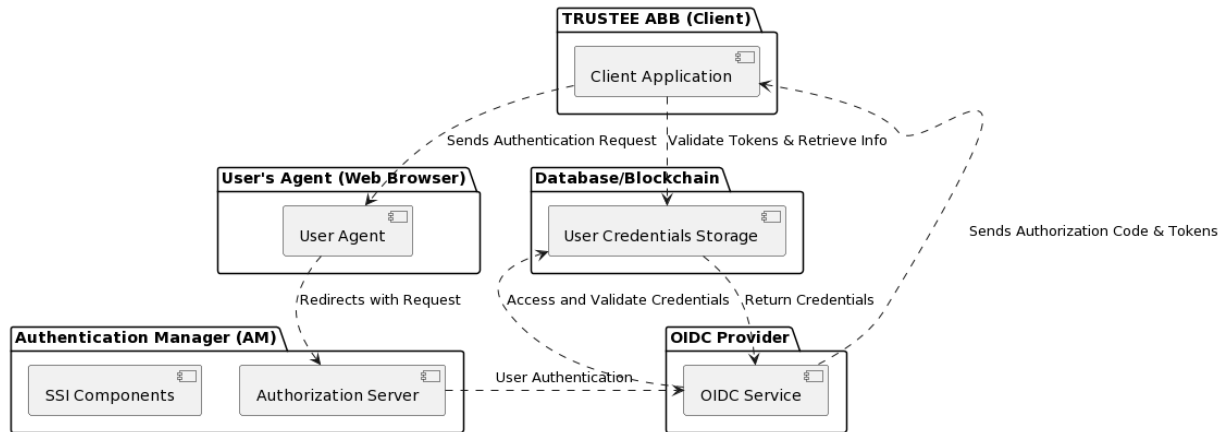


Figure 87 OIDC Authentication Process

The UML component diagram in Figure 87 for the OIDC Authentication Process presents a structural view of the system, focusing on the components and their interactions. An elaboration of each component and subcomponent, along with their roles and interconnections is as follows:

- **TRUSTEE ABB (Client) - [Client Application]:** This component represents the client-side application that initiates the authentication process. It prepares an authentication request containing parameters like `client_id`, `scope`, `response_type`, and `nonce`. The Client Application is responsible for sending the authentication request to the User's Agent and for receiving and processing the tokens from the OIDC Provider.
- **User's Agent (Web Browser) - [User Agent]:** The User Agent is typically a web browser used by the end-user. It acts as an intermediary between the Client Application and the Authentication Manager. The User Agent is responsible for redirecting the user with the authentication request to the Authorization Server and for redirecting the user back to the Client Application with the authorization code.
- **Authentication Manager (AM):** This is a complex component consisting of the Authorization Server and SSI Components.
 - **[Authorization Server]:** This subcomponent is crucial for authenticating the user and obtaining end-user consent. It interacts with the OIDC Provider to authenticate the user and generate the authorization code and tokens.
 - **[SSI Components]:** These components (like Aries Controller and Agent) are involved in handling Self-Sovereign Identity (SSI) aspects, like managing blockchain interactions for user credentials.
- **OIDC Provider - [OIDC Service]:** The OIDC Provider is a core component of the Authentication Manager. It handles the OIDC protocol, particularly the authorization and token exchange. The OIDC Service sends the authorization code and tokens back to the Client Application and communicates with the Database/Blockchain for validating user credentials and generating tokens.
- **Database/Blockchain - [User Credentials Storage]:** This component is responsible for storing and managing user credentials and other authentication-related data. It interacts with the OIDC Service to validate credentials and provide the necessary information for token generation.

8.5.2.2. REGISTERING A USER

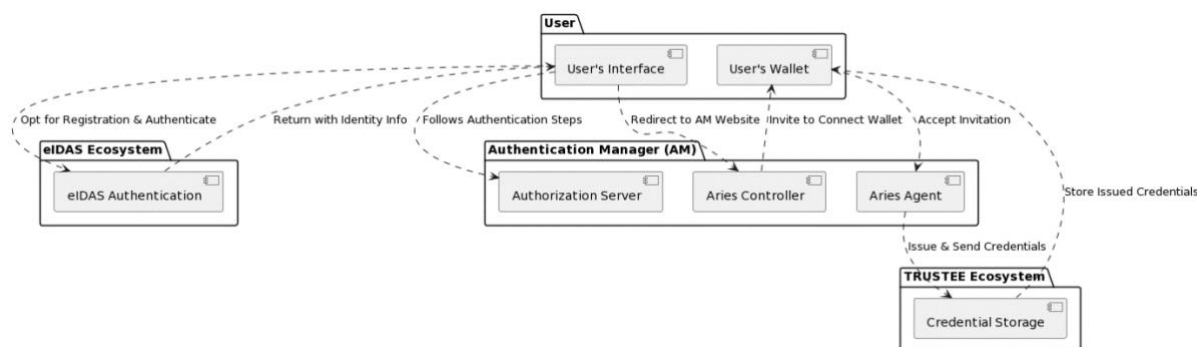


Figure 88 Secure and verifiable user identity within the TRUSTEE ecosystem

The UML component diagram of Figure 88 depicts the process of user registration and SSI credential issuance. A detailed explanation of each component, subcomponent, and their roles is as follows:

- User:
 - User's Interface: This component represents the platform or interface through which the user interacts with the Authentication Manager and the eIDAS ecosystem. It facilitates the user's journey from starting the authentication process to choosing to register and acquire SSI credentials. This interface is where the user will select options for registration and receive directions for further steps.
 - User's Wallet: A digital wallet on the user's device, used to store SSI credentials. Once the credentials are issued by the TRUSTEE ecosystem, they are stored in this wallet. The wallet is also used to accept invitations from the Aries Controller for credential issuance.
- eIDAS Ecosystem - [eIDAS Authentication]: A key component responsible for the secure and verified authentication of the user's identity. The user, through their interface, engages with this component using an eIDAS-compatible identity card or method. It verifies the user's identity and returns them to the Authentication Manager's website with their authenticated identity information.
- Authentication Manager (AM):
 - Authorization Server: This subcomponent of AM is involved in the initial authentication steps as per the OIDC protocol. It interacts with the user's interface during the initial phase of the authentication process.
 - Aries Controller: A crucial part of AM, the Aries Controller takes over once the user decides to register for SSI credentials. It invites the user to connect their digital wallet and initiates the credential issuance process by instructing the Aries Agent.
 - Aries Agent: This subcomponent processes the credential issuance. Upon receiving instructions from the Aries Controller and acceptance from the user's wallet, the Aries Agent issues the SSI credentials and sends them to the user's wallet.
- TRUSTEE Ecosystem - [Credential Storage]: Represents the system where the issued credentials are securely stored and managed. Once the Aries Agent issues the credentials, they are sent to this component for storage. The user's wallet interacts with this component to store and manage the newly issued credentials.

8.5.3. TECHNICAL DETAILS

In this section, technical information will be provided for each component. The technical details will be provided in the form of a table listing the most important technical information of each Authentication Manager component or subcomponent. This information will change as the implementation's maturity change. To allow for easy deployment of the Authentication Manager, to avoid software conflicts and minimise software requirements and to minimise the resources required, all the components of the Authentication Manager are dockerised. This means the Authentication Manager can be deployed using Docker compose and thus, the only requirement for the host system is the Docker server.

Component: Authentication Agent:

Development	
Type (Software/Hardware)	Software
Operating Systems	ANY
Database and Databases tools	SQL
GUI	Bootstrap, HTML5, CSS3, Vanilla JavaScript
Programming Language	Python, HTML, CSS, JavaScript
Installation and Deployment	
Software Requirements	Python3, Flask
Hardware Requirements	-
Containerization	Docker, Docker Compose
Communications	HTTP(s)

Subcomponent: Authentication Agent > OIDC Provider:

Development	
Type (Software/Hardware)	Software

Operating Systems	ANY
Database and Databases tools	SQL
GUI	Bootstrap, HTML5, CSS3, Vanilla JavaScript
Programming Language	Python, HTML, CSS, JavaScript
Installation and Deployment	
Software Requirements	Python3, Flask, Authlib
Hardware Requirements	-
Containerization	Docker, Docker Compose
Communications	HTTP(s), OIDC, Auth2.0

Subcomponent: Authentication Agent > Aries Controller:

Development	
Type (Software/Hardware)	Software
Operating Systems	ANY
Database and Databases tools	SQL
GUI	Bootstrap, HTML5, CSS3, Vanilla JavaScript
Programming Language	Python, HTML, CSS, JavaScript
Installation and Deployment	
Software Requirements	Python3, Flask, Python Requests

Hardware Requirements	-
Containerization	Docker, Docker Compose
Communications	HTTP(s)

Subcomponent: Authentication Agent > eIDAS Bridge:

Development	
Type (Software/Hardware)	Software
Operating Systems	ANY
Database and Databases tools	-
GUI	-
Programming Language	Python
Installation and Deployment	
Software Requirements	Python3, Flask
Hardware Requirements	-
Containerization	Docker, Docker Compose
Communications	HTTP(s), SAML

Component: Aries Agent:

Development

Type (Software/Hardware)	Software
Operating Systems	ANY
Database and Databases tools	-
GUI	-
Programming Language	Python
Installation and Deployment	
Software Requirements	ACA-Py, indy-sdk, libindy
Hardware Requirements	-
Containerization	Docker, Docker Compose
Communications	HTTP(s), DIDComm

Component: SSI blockchain:

Development	
Type (Software/Hardware)	Software
Operating Systems	ANY
Database and Databases tools	-
GUI	-
Programming Language	-
Installation and Deployment	

Software Requirements	VON Network, Hyperledger Indy
Hardware Requirements	-
Containerization	Docker, Docker Compose
Communications	HTTP(s), DID

8.6. VALIDATION

A lot of work was done on the Authentication Manager since the beginning of the TRUSTEE project. A solid architecture was designed to facilitate the needs of the project and the main functionalities of the Authentication Manager were implemented.

8.6.1. VALIDATION METHODOLOGY

The Authentication Manager (AM) validation methodology presented below establishes a framework for validating the AM building block. Since not all the internal components of the Authentication Manager are mature enough and not all the external ABBs are mature enough for integration, it is not yet possible to conduct a complete validation. Thus, this methodology will be utilized further in the upcoming deliverables and on each iteration validate the components that are at the appropriate level of development.

- **Validation Objectives** – Clearly define the objectives for AM validation, ensuring adherence to functional requirements, performance benchmarks, security standards, and usability criteria.
- **Validation Criteria** – Establish specific criteria, including accuracy, response time, error rates, scalability, compliance, and user satisfaction, to gauge the success of the validation process for AM.
- **Test Plans and Execution** – Develop comprehensive test plans covering various aspects such as functionality, performance, security, and usability, based on AM's use cases. Execute the test cases diligently, carefully assessing the results and taking note of any issues or discrepancies for prompt resolution.
- **Integration and Testing** – Carry out integration testing to evaluate how Authentication Manager interacts with external ABBs. Present functionalities and communicate with other ABB stakeholders ensuring the efficient communication and collaboration between components.
- **Evaluation and Iterative Improvement** – Analyze the test results, compare them against the validation criteria, and identify areas that need improvement. Address any issues, bugs, or performance gaps through iterative development and additional testing for continual enhancement.
- **Documentation and Review** – Document the AM validation process, including test cases, and modifications made. Seek review and approval from relevant stakeholders to ensure alignment with project objectives.

8.6.2. VALIDATION & UNIT TESTING

The evaluation and validation of the core functionalities of the Authentication Agent is done by conducting tests where the related user scenarios are emulated and tested from start to end. This enables us to identify any fault affecting the authentication or registration process, but at the same time, depending on the internal component in

question may make the development of the components harder, as all the major functionalities will have to be implemented before being able to test them.

To facilitate the timely development independently of the available resources (e.g. servers, domains, etc.), only for the development some security features will temporary be disabled to bypass limitations and allow testing and debugging. For example, during testing the requirement of using HTTPS was disabled so that the implementation could be evaluated without the need of TLS certificates and official TRUSTEE domain names configured on our development servers and without the need to wait for the availability of these resources.

During the initial stage of the development, where different ABBs will have reach different maturity levels, we opted for emulating their side of using the Authentication Manager so that we can test all its features independently.

8.7. NEXT STEPS

The next steps for the development, integration and validation of the AM ABB along with the TRUSTEE platform include the following:

1. **Enhanced Integration of SSI Components with OIDC Protocol:** Focus on integrating the SSI components (like Aries Controller and Agent) more seamlessly with the OIDC protocol. This includes ensuring that the SSI credentials are properly utilized during the OIDC authentication process and that the data flow between the Aries components and the OIDC Provider is optimized.
2. **Integration with the DA ABB (Dashboard):** The interface with the DA should facilitate easy registration and management of OIDC applications. For users, the interface should provide a clear and straightforward journey through authentication, eIDAS-based verification, and credential issuance processes.
3. **Implementation of Advanced Security Measures:** Implement additional security protocols and encryption methods to protect sensitive user data and credentials. This includes securing the communication channels between different components of the system and ensuring compliance with data protection regulations.
4. **Comprehensive Testing and Validation:** Conduct thorough testing of the entire ecosystem. This involves unit testing of individual components, integration testing to ensure smooth interaction between different parts of the system, and user acceptance testing to validate the end-to-end process from the user's perspective. Special attention should be given to testing the resilience of the system against various security threats.
5. **Feedback Loop and Iterative Improvement:** Establish a feedback mechanism to gather insights from both technical stakeholders and end-users. Use this feedback to iterate and improve the system continuously. This includes refining the user interface based on user experience, optimizing the performance of the system, and updating security measures in response to emerging threats and challenges.

8.8. CONCLUSIONS

The authentication manager will allow the authentication and onboarding of users, allowing all the ABBs inside the TRUSTEE ecosystem to authenticate users and enforce an access control limiting access to the TRUSTEE services only to authorised identities and allow actions to be made only from authorised parties. To achieve these, authentication manager will leverage the SSI technology and the eIDAS services.

9. CONCLUSION

This document presents the initial version of the TRUSTEE ABBs developed under WP3:

- HEDF
- AIMAAS
- TAI-SDF
- ATR
- KR
- AM

Each of the above-mentioned ABBs was described in a respective chapter in this document. For each ABB the needs and objectives have been discussed. Some basic concepts needed to understand the ABB design and implementation have also been included as Background. Then, the architecture design and implementation details have been exposed according to each ABB maturity. Finally, some validation tests have been proposed (and presented when available). The next steps for the following versions of each ABB have been presented, as well as the main conclusions of the current implementation for each ABB.

Throughout the implementation process and the technical development activities and investigation performed under WP3, as well as following the project's progress, additional technical requirements may emerge. As explained in the introduction section of this document, WP3 will provide feedback to WP2 under which requirements and architecture details are identified and reported, therefore additional technical requirements stemming from the implementation process will be reported in the upcoming D2.2 – “Live doc conceptualisation, use cases and system architecture V2”. The description of the presented ABBs will be revised, enhanced, and complemented as the project progresses and the maturity of the ABBs improves. The updated details will be gathered in future versions of the deliverable, namely D3.2 and D3.3.

10. REFERENCES

- D2.1 TRUSTEE D2.1. TRUSTEE D2.1 Live doc conceptualisation, use cases and system architecture, 2023.
- D4.1 TRUSTEE D4.1 Live document, TRUSTEE Automation, OPEN CONTINUUM & SSI Homomorphic capable framework
- Ahmad 2022 Ahmad, S., et al. (2022). Data Fusion Techniques: A Comprehensive Review. *IEEE Access*, 10, 2968-2991.
- Avanti 2017 Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. "Learning Important Features Through Propagating Activation Differences". In: arXiv preprint arXiv:1704.02685 (2017).
- Avram 2023 Avram, A. (n.d.). *Docker: Automated and Consistent Software Deployments*. Retrieved 10 24, 2023, from <http://www.infoq.com/news/2013/03/Docker>
- Benaissa 2021 A. Benaissa, B. Retiat, B. Cebere, and A. E. Belfedhal, "Tenseal: A library for encrypted tensor operations using homomorphic encryption," 2021
- Benet 2014 Benet, J. (2014). IPFS - Content Addressed, Versioned, P2P File System. *arXiv: Networking and Internet Architecture*. Retrieved 10 24, 2023, from <https://arxiv.org/abs/1407.3561>
- Benhamouda 2018 Benhamouda, F., Halevi, S., & Halevi, T. (2018). *Supporting Private Data on Hyperledger Fabric with Secure Multiparty Computation*. Retrieved 10 24, 2023, from <https://ieeexplore.ieee.org/document/8360354>
- Candelieri 2022 Candelieri, A., et al. (2022). Gaia-X: A federated architecture for trustworthy data sharing and collaboration. *Journal of Parallel and Distributed Computing*, 169, 65-80.
- Chen 2021 Chen, Y., Li, J., & Carin, L. (2021). Adversarial attacks and defenses in time series classification: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 32(4), 1423-1440

- Chen 2022 Chen, Y., Li, J., & Carin, L. (2022). Adversarial attacks and defenses in graph neural networks: A survey. arXiv preprint arXiv:2201.05005
- Chui 2018 Chui M, Malhotra S (2018) AI adoption advances, but foundational barriers remain. <https://www.mckinsey.com/featured-insights/artificial-intelligence/ai-adoption-advances-but-foundational-barriers-remain>. Accessed 13 July 2020
- Crawford 2019 Crawford K (2019) Halt the use of facial-recognition technology until it is regulated. *Nature* 572:565. <https://doi.org/10.1038/d41586-019-02514-7>
- Decan 2018 Alexandre Decan, Tom Mens, and Eleni Constantinou. On the impact of security vulnerabilities in the npm package dependency network. In *Proceedings of the 15th International Conference on Mining Software Repositories*, pages 181–191, 2018
- Decan 2019 Alexandre Decan, Tom Mens, and Philippe Grosjean. An empirical comparison of dependency network evolution in seven software packaging ecosystems. *Empirical Software Engineering*, 24(1):381–416, 2019
- Dosilove 2018 F. K. Dosilović, M. Brčić, and N. Hlupić, “Explainable artificial intelligence: A survey,” in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2018, pp. 0210–0215.
- Dwork 2006 Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006, March). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference* (pp. 265-284). Springer, Berlin, Heidelberg.
- Dutta 2018 Dutta S (2018) An overview on the evolution and adoption of deep learning applications used in the industry. *Wires Data Min Knowl Discov* 8:1–12. <https://doi.org/10.1002/widm.1257>
- Elshaw 2018 Elshaw R, Sakr S, Talia D, Trunfio P (2018) Big data systems meet machine learning challenges: towards big data science as a service. *Big Data Res* 14:1–11. <https://doi.org/10.1016/j.bdr.2018.04.004>
- Ge 2022 Ge, H., et al. (2022). Exploiting scientific publications for data fusion: A systematic review. *Information Fusion*, 78, 44-60.

- Geman 1995 D. Geman, C. Yang, “Nonlinear image recovery with half-quadratic regularization”, IEEE transactions on Image Processing, no. 4, vol. 7, pp. 932-946, 1995.
- Guizzardi 2023 Guizzardi, G., & Guarino, N. (2023). Semantics, Ontology and Explanation. arXiv preprint arXiv:2304.11124.
- He, J 2021 He, J., et al. (2021). Distributed Indexing Framework for Large-Scale Data Fusion Systems. Future Generation Computer Systems, 125, 367-381.
- Huang 2022 Huang, Y., et al. (2022). Multimodal Data Fusion: A Comprehensive Review of Recent Advances. IEEE Transactions on Cybernetics, 52(3), 1195-1210.
- Imtiaz 2021 Nasif Imtiaz, Seaver Thorn, Laurie A. Williams: A comparative study of vulnerability reporting by software composition analysis tools. ESEM 2021: 5:1-5:11
- IPFS Paper n.d. IPFS Paper. (n.d.). Retrieved 10 24, 2023, from <https://github.com/ipfs/papers/raw/master/ipfs-cap2pfs/ipfs-p2p-file-system.pdf>
- Khan 2022 M. Khan and J. Vice, “Toward Accountable and Explainable Artificial Intelligence Part one: Theory and Examples,” 2 2022. [Online]. Available: <https://www.techrxiv.org/articles/preprint/Toward Accountable and Explainable Artificial Intelligence Part one Theory and Examples/19102085>
- Kim 2018 M. Kim, O. Reingold, and G. Rothblum, “Fairness through computationally-bounded awareness,” in Advances in Neural Information Processing Systems, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018. [Online]. Available: <https://proceedings.neurips.cc/paper/2018/file/c8dfece5cc68249206e4690fc4737a8d-Paper.pdf>
- Konečný 2016 Konečný, J., McMahan, H. B., Ramage, D., & Richtárik, P. (2016). Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*.
- Li 2022 Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou, Trustworthy AI: From Principles to Practices, 2022 Association for Computing Machinery

- Li, Y. 2022 Li, Y., et al. (2022). Semantic-Aware Indexing for Heterogeneous Data Fusion: State of the Art and Future Directions. *Information Fusion*, 80, 251-269.
- Liu 2022 Haochen Liu, Yiqi Wang, Wenqi Fan, Xiaorui Liu, Yaxin Li, Shaili Jain, Yunhao Liu, Anil Jain, and Jiliang Tang. 2022. *Trustworthy AI: A Computational Perspective*. *ACM Trans. Intell. Syst. Technol.* 14, 1, Article 4 (February 2023), 59 pages. <https://doi.org/10.1145/3546872>
- Liu 2023 Liu, Y., Li, Y., & Wang, X. (2023). Adversarial attacks and defenses in computer vision: A survey. *Journal of Visual Communication and Image Representation*, 87, 103174
- Liu W. 2022 Liu, W., et al. (2022). A Unified Indexing Framework for Heterogeneous Data Fusion. *IEEE Transactions on Knowledge and Data Engineering*, 34(1), 221-235.
- Lundberg 2017 Lundberg, Scott M., and Su-In Lee. “A unified approach to interpreting model predictions.” *Advances in Neural Information Processing Systems* (2017).
- Machanavajjhala 2007 Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1), 3-es.
- Madani 2018 Madani A, Arnaout R, Mofrad M, Arnaout R (2018) Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digit Med*. <https://doi.org/10.1038/s41746-017-0013-1>
- Mahmoud 2023 Mahmoud Alfadel, Diego Elias Costa, Emad Shihab: Empirical analysis of security vulnerabilities in Python packages. *Empir. Softw. Eng.* 28(3): 59 (2023)
- Manoj 2020 Manoj, M. K., & Krishnan, S. S. (2020). *Decentralizing Privacy Using Blockchain to Protect Private Data and Challenges With IPFS*. Retrieved 10 24, 2023, from <https://iglobal.com/chapter/decentralizing-privacy-using-blockchain-to-protect-private-data-and-challenges-with-ipfs/238369>
- Mittal 2021 Mittal, P., et al. (2021). Data Acquisition Methods: A Comprehensive Review and Analysis. *IEEE Access*, 9, 19653-19674.

- Monga 2021 V. Monga, Y. Li, and Y. C. Eldar, “Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing,” *IEEE Signal Processing Magazine*, vol. 38, no. 2, pp. 18–44, 2021.
- Nasir 2018 Nasir, Q., Qasse, I. A., Talib, M. A., & Nassif, A. B. (2018). Performance Analysis of Hyperledger Fabric Platforms. *Security and Communication Networks*, 2018, 1-14. Retrieved 10 24, 2023, from <https://hindawi.com/journals/scn/2018/3976093>
- Nicolae 2018 Maria-Irina Nicolae, Mathieu Sinn, Tran Ngoc Minh, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Ian M. Molloy, Benjamin Edwards: Adversarial Robustness Toolbox v0.2.2. CoRR abs/1807.01069 (2018)
- Papernot 2016 Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., & Talwar, K. (2016). Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*.
- Phillips 2021 P. J. Phillips, C. Hahn, P. Fontana, A. Yates, K. K. Greene, D. Broniatowski, and M. A. Przybocki, “Four principles of explainable artificial intelligence,” 2021-09-29 04:09:00 2021. [Online]. Available: <https://tsapps.nist.gov/publication/getpdf.cfm?pubid=933399>
- Rai 2019 Rai A, Constantinides P, Sarker S (2019) Next-generation digital platforms: toward human-AI hybrids. *MIS Q* 43:iii–ix
- Ribeiro 2016 Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. “Why should I trust you?: Explaining the predictions of any classifier.” *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM (2016).
- Romero 2019 Romero F, Li Q, Yadwadkar NJ, Kozyrakis C (2019) INFaaS: managed & model-less inference serving. <https://arxiv.org/pdf/1905.13348.pdf>. Accessed 31 March 2020
- Salimitari 2019 Salimitari, M., Joneidi, M., & Chatterjee, M. (2019). An Outlier-aware Consensus Protocol for Blockchain-based IoT Networks Using Hyperledger Fabric. *arXiv: Distributed, Parallel, and Cluster Computing*. Retrieved 10 24, 2023, from <https://arxiv.org/pdf/1906.08177>
- Sayeed 2019 Sayeed, S., & Marco-Gisbert, H. (2019). Assessing blockchain consensus and security mechanisms against the 51% attack. *Applied Sciences*, 9(9), 1788. Retrieved

- 10 24, 2023, from <https://research-portal.uws.ac.uk/en/publications/assessing-blockchain-consensus-and-security-mechanisms-against-th>
- Shan 2020 T. Shan, J. Wang, F. Chen, P. Szenher, B. Englot, “Simulation-based lidar super-resolution for ground vehicles”, *Robotics and Autonomous Systems*, 134, 103647, 2020.
- Sharma 2018 Sharma, A., Schuhknecht, F. M., Agrawal, D., & Dittrich, J. (2018). How to Databasify a Blockchain: the Case of Hyperledger Fabric. *arXiv: Distributed, Parallel, and Cluster Computing*. Retrieved 10 24, 2023, from <https://arxiv.org/abs/1810.13177>
- Singh 2021 Singh, S., et al. (2021). Missing Data Imputation Techniques in the Context of Data Fusion: A Review. *Information Fusion*, 73, 77-95.
- Soelman 2019 Soelman, M. (2019). *Hyperledger Fabric: A study of endorsement policies for supply chains*. Retrieved 10 24, 2023, from <http://fse.studenttheses.ub.rug.nl/20018>
- Steichen 2018 Steichen, M., Fiz, B., Norvill, R., Shbair, W. M., & State, R. (2018). *Blockchain-Based, Decentralized Access Control for IPFS*. Retrieved 10 24, 2023, from <https://orbi.lu.uni.lu/handle/10993/36641>
- Thiebes 2020 Thiebes S, Lins S, Sunyaev A (2020) Trustworthy artificial intelligence. *Electron Mark*. <https://doi.org/10.1007/s12525-020-00441-4>
- Tindall 2016 Tindall, L. (2016). *IPFS, the Interplanetary File System*. Retrieved 10 24, 2023, from <https://nora.codes/post/ipfs-the-interplanetary-file-system>
- Wang 2018 Wang W, Gao J, Zhang M, Wang S, Chen G, Ng TK, Ooi BC, Shao J, Reyad M (2018) Rafiki: machine learning as an analytics service system. *Proc VLDB Endow* 12:128–140. <https://doi.org/10.14778/3282495.3282499>
- Wang 2018 Wang, M., Duan, M., & Zhu, J. (2018). *Research on the Security Criteria of Hash Functions in the Blockchain*. Retrieved 10 24, 2023, from <https://dl.acm.org/citation.cfm?id=3205238>
- Wang 2020 Wang, Y., Li, Y., & Wang, X. (2020). Adversarial attacks and defenses in reinforcement learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9), 3405-3424

- Wang 2021 Wang, Y., Li, Y., & Wang, X. (2021). Adversarial attacks and defenses in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology*, 12(2), 1-28.
- Weller 2017 A. Weller, “Challenges for transparency,” *CoRR*, vol. abs/1708.01870, 2017. [Online]. Available: <http://arxiv.org/abs/1708.01870>
- Yakubov 2018 Yakubov, A., Shbair, W. M., Wallbom, A., Sanda, D., & State, R. (2018). *A Blockchain-Based PKI Management Framework*. Retrieved 10 24, 2023, from <https://ieeexplore.ieee.org/document/8406325>
- Yao 1986 Yao, A. C. C. (1986, October). How to generate and exchange secrets. In *27th Annual Symposium on Foundations of Computer Science (sfcs 1986)* (pp. 162-167). IEEE.
- Zhang 2017 K. Zhang, W. Zuo, S. Gu, L. Zhang, “Learning deep CNN denoiser prior for image restoration”, In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.
- Zhang 2020 Zhang, Y., Xu, Y., & Wang, X. (2020). Adversarial attacks and defenses in deep learning. *IEEE Access*, 8, 91937-91952.
- Zhang 2021 Zhang, Y., Xu, Y., & Wang, X. (2021). Adversarial attacks and defenses in federated learning: A survey. *IEEE Transactions on Information Forensics and Security*, 16, 1-20
- Zhang L 2021 Zhang, L., et al. (2021). A Unified Indexing Framework for Data Fusion in Internet of Things. *IEEE Access*, 9, 96570-96583.
- Zhang Z 2022 Zhang, Z., Zhao, J., Huang, C., & Li, L. (2022). Learning visual semantic map-matching for loosely multi-sensor fusion localization of autonomous vehicles. *IEEE Transactions on Intelligent Vehicles*, 8(1), 358-367.

11. ANNEX 1: GLOSSARY OF TERMS

This section contains the more important terms for the understanding of the deliverable. More details can be found in D2.1 glossary definition (TRUSTEE consortium, 2023).

Architecture Building Block (ABB) - a constituent of the architecture model that describes a single aspect of the overall model. An Architecture Building Block describes generic characteristics and functionalities. Architecture Building Blocks are used to describe reference architectures, solution architecture templates or solution architectures of specific solutions. (TOGAF) (see **Building Block**)

Artificial intelligence (AI) – Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. The AI system can generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions. As a scientific discipline, AI includes several approaches and techniques, such as machine learning approaches (including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning), machine reasoning (which includes, in particular, planning, scheduling, search, and optimization), logic- and knowledge-based approaches (including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems), statistical approaches, Bayesian estimation and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber- physical systems). [White Paper On Artificial Intelligence - A European approach to excellence and trust. European Commission. COM(2020) 65 final; [HLEG AI, Assessment List for Trustworthy AI \(ALTAI\)](#); Proposal of the EU Artificial Intelligence Act (Brussels, 21.4.2021 COM(2021) 206 final 2021/0106 (COD)), Article 2, point 1.]

Asymmetric cryptography - cryptography that uses two separate keys to exchange data, one to encrypt or digitally sign the data and one for decrypting the data or verifying the digital signature. Also known as public key cryptography. [[NIST SP 800-77 Rev. 1](#) under Asymmetric Cryptography]

Attribute - inherent property or characteristic of an entity that can be distinguished quantitatively or qualitatively by human or automated means [ISO/IEC 25000:2014]

Automatic/Automated/Automation - pertaining to a process or system that, under specified conditions, functions without human intervention. [[ISO/IEC DIS 22989\(en\). Terms related to Artificial Intelligence](#)].

Building Block - a (potentially re-usable) component of enterprise capability that can be combined with other building blocks to deliver architectures and solutions. Building blocks can be defined at various levels of detail, depending on what stage of architecture development has been reached. For instance, at an early stage, a building block can simply consist of a name or an outline description. Later, a building block may be decomposed into multiple supporting building blocks and may be accompanied by a full specification. Building blocks can relate to "architectures" or "solutions". According to The Open Group, building blocks have several generic characteristics: see **Architecture Building Block**, **Solution Building Block**. According to The Open Group, building blocks have several generic characteristics: A building block is a package of functionality defined to meet the business needs across an organization, A building block has published interfaces to access the functionality, A building block may have multiple implementations but with different inter-dependent “building blocks.” A building block may interoperate with other, interdependent building blocks, A good building block considers implementation and usage

and evolves to exploit technology and standards, It may be assembled from other building blocks, It may be a subassembly of other building blocks, Ideally, a building block is re-usable and replaceable, and well-specified.

Certificate Authority (CA) - a trusted entity that issues and revokes public key certificates. [NIST SP 1800-16B under Certificate Authority from NISTIR 8149]

Ciphertext (CT)— the unintelligible and seeming random form of data that is produced by the cryptographic function of encryption. Ciphertext is produced by a symmetric algorithm when a data set is transformed by the encryption process using a selected key. Ciphertext can be converted back into its original form (i.e. plain text) by performing the decryption process using the same symmetric encryption algorithm and the key used during the encryption process. (Also known as a cryptogram.) [The A to Z of Cybersecurity Glossary by [Simplicity](#)]

Confidentiality - the property that information is not made available or disclosed to unauthorized individuals, entities, or processes. The purpose of the confidentiality service is to protect information from unauthorized disclosure. [ITU-TX.800] **Cross-border processing** – means either processing of personal data which takes place in the context of the activities of an organisation in more than one Member State where that organisation is established in more than one Member State; or processing of personal data which takes place in the context of the activities of an organisation's single establishment but where that processing substantially affects or is likely to substantially affect data subjects in more than one Member State. [GDPR, Article 4(23)]

Consent - consent of the data subject means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her. [Article 4(11) of the GDPR]

Control - a measure that is modifying risk. Controls include any process, policy, device, practice, or other actions that modify risk. (ISO/IEC 27000: 2018)

Cross-border processing – means either processing of personal data which takes place in the context of the activities of an organisation in more than one Member State where that organisation is established in more than one Member State; or processing of personal data which takes place in the context of the activities of an organisation's single establishment but where that processing substantially affects or is likely to substantially affect data subjects in more than one Member State. [GDPR, Article 4(23)]

Data - a representation of facts, concepts, or instructions in a manner suitable for communication, interpretation, or processing by humans or by automatic means [IEEE Glossary]. A digital representation of acts, facts or information and any compilation of such acts, facts or information, including in the form of sound, visual or audiovisual recording. [DGA, [Article 2\(1\)](#)]

Data Confidentiality - the property that information is not made available or disclosed to unauthorized individuals, entities, or processes. (ISO/IEC PDTR 13335-1). Confidentiality in a general sense refers to the duty not to share information with persons who are not qualified to receive that information (see Article 5(f) of Regulation (EU) 2016/679) [[EDPS Data protection Glossary](#)].

Data holder – legal person, including public sector bodies and international organisations, or a natural person who is not a data subject with respect to the specific data in question, which, in accordance with applicable law, has the right to grant access to or to share certain personal data or non-personal data [Article 2(8)DGA]; means a legal or natural person, who has accessed data from the connected product or has generated data during the provision of a related service and who has the contractually agreed right to use such data, and the obligation, in accordance with this Regulation, applicable Union law or national legislation implementing Union law to make available certain data to the user or a data recipient [Data Act, Article 2(6), version after amendments of the EP, available at: [Texts adopted - Data Act - Tuesday, 14 March 2023 \(europa.eu\)](#)]

Data management - the activities of defining, creating, storing, maintaining and providing access to data and associated processes in one or more information systems. [b-ISO/IEC TR 10032:2003]

Data processing - the ability to handle data as input and apply different treatments that might modify the data or combine it without modifying it with other data in order to produce an output that is useful for a given application or service in the data lifecycle. In terms of the GDPR: *Any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction* [Article 4(2) of the GDPR]. Related definition – **restriction of processing** (see below).

Data processor - see **Processor**.

Data protection - means the implementation of appropriate administrative, legal, technical or physical means to guard against unauthorized intentional or accidental disclosure, modification, or destruction of data.

Data provider - organization which produces or collects and further shares data or metadata. [OECD]

Data recipient – in the context of personal data protection: means a natural or legal person, public authority, agency or another body, to which the personal data are disclosed, whether a third party or not. However, public authorities which may receive personal data in the framework of a particular inquiry in accordance with Union or Member State law shall not be regarded as recipients; the processing of those data by those public authorities shall be in compliance with the applicable data protection rules according to the purposes of the processing. [Article 4(9) of the GDPR]. In the context of Data Act: means a legal or natural person other than the user of a connected product or related service, to whom a data holder makes available data accessed from a connected product or generated during the provision of a related service following an explicit request by the user or in accordance with a legal obligation under Union law or national legislation implementing Union law [Data Act, Article 2(7), version after amendments of the EP, available at: [Texts adopted - Data Act - Tuesday, 14 March 2023 \(europa.eu\)](#)]

Data trust - includes actions taken to safeguard security, privacy, and quality of data and enhance data trustworthiness.

Data use - after the data visualization and analysis, the data use is the process that application run to either understand the monitored phenomena by analyzing the past and monitor the present or predict the future and build new decisions.

Data user - a natural or legal person who has lawful access to certain personal or non-personal data and has the right, including under the GDPR in the case of personal data, to use that data for commercial or non-commercial purposes. [DGA, Article 2(9)]

Developer - individual or organization that performs development activities (including requirements analysis, design, testing through acceptance) during the system or software life-cycle process [: ISO/IEC 25000:201]

FAIR Data principles - the FAIR Data Principles are a set of guiding principles in order to make data findable, accessible, interoperable and reusable.

Homomorphic Encryption (Homomorphic Cryptography) – encryption schemes allowing certain mathematical operations to be performed directly on ciphertexts, without prior decryption. Homomorphic encryption can be a powerful tool for leveraging multi-party computations, by enabling a participant to compute functions on values while keeping the values hidden.

Legal Requirement - a responsibility placed on an entity, including statutory or regulatory obligations which an entity is required to fulfil, to ensure legal compliance of its actions.

Metadata - data collected on any activity of a natural or legal person for the purposes of the provision of a data sharing service, including the date, time and geolocation data, duration of activity, connections to other natural or legal persons established by the person who uses the service. [DGA COM (2020) 767 FINAL]

Non-disclosure agreement – a legal contract between two parties or among more parties stating the use and sharing of data among the parties for certain purposes but restricts access to.

Personal data - means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person [Regulation (EU) 2016/679 (General Data Protection Regulation), Article 4 (1)]. Data which relates to a living individual who can be identified: a) from those data; or b) from those data and other information which is in possession of, or is likely to come into the possession of, the data controller, and includes any expression of opinion about the individual and any indication of the intentions of the data controller or any other person in respect of the individual. [b-PAS 185:2017]

Personally Identifiable Information (PII) - personally Identifiable Information is defined as any information about an individual maintained by an [entity], including any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date and place of birth, and any other information that is linked or linkable to an individual([NIST10] p. 2-1). PII is a widely used term for personal data in the domain of information security. Within this document, PII is used in relation to information security

Privacy Impact Assessment (PIA) – in other words Data Protection Impact Assessment (DPIA) – describes a process designed to identify risks arising out of the processing of personal data and to minimise these risks as far and as early as possible. DPIAs are important tools for negating risk, and for demonstrating compliance with the GDPR. Required by Article 35 of the GDPR where data processing “is likely to result in a high risk to the rights and freedoms of natural persons”.

Sensitive data - any information that is protected against unwarranted disclosure. Protection of data may be required for legal or ethical reasons, for issues pertaining to personal privacy, or for proprietary considerations. [CSC]

Sensitive Personal Data / Special category of personal data – Personal data which are, by their nature, particularly sensitive in relation to fundamental rights and freedoms merit specific protection as the context of their processing could create significant risks to the fundamental rights and freedoms. In the GDPR sensitive personal data are considered as “special categories of personal data”. The following personal data is considered ‘special category of personal data’ and is subject to specific processing conditions stemming in particular from the GDPR: personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, genetic data, biometric data processed for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation. [Article 4(13), 4(14), 4(15), 9(1) and Recitals 51 to 56 of the GDPR, https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive_en]

12. ANNEX 2: LEGAL REQUIREMENTS

The following sections detail the preliminary analysis of legal requirements presented in the D2.1 (TRUSTEE consortium, 2023) related to each component. The ID of the respective legal requirement is used as a reference to D2.1, while the requirement's description is also provided alongside whether the ABBs comply with the requirement and how. The approach presented below for addressing each legal requirement may produce additional technical, functional and non-functional, requirements that will be further specified in upcoming versions of this deliverable, namely D3.2 and D3.3, as well as in D2.2.

12.1. HEDF

The following sections detail the preliminary analysis of legal requirements presented in the D2.1 (TRUSTEE consortium, 2023) related to each component. The ID of the respective legal requirement is used as a reference to D2.1, while the requirement's description is also provided alongside whether the ABBs comply with the requirement and how. The approach presented below for addressing each legal requirement may produce additional technical, functional and non-functional, requirements that will be further specified in upcoming versions of this deliverable, namely D3.2 and D3.3, as well as in D2.2.

12.1. HEDF

The following table shows the legal requirements as addressed by the HEDF component.

Req id	Req. Description	Compliance	How is addressed
LEG-USR-003	The Data Provider / Consumer should share the data in pseudonymised, or, where possible, in anonymised form, to ensure data minimisation and purpose limitation	Yes	No raw data of any form is directly accessed by HEDF
LEG-USR-016	When a data subject exercises its right of erasure and/or restriction to the processing, Data Provider must ensure to remove the data from TRUSTEE / stop sharing the data of data subjects when requested by a data subject, required by the data protection laws and when there is no legal basis for further sharing and processing of the data.	Partially	This is under investigation, but is a requirement that will be fulfilled in the future implementations.
LEG-USR-018	Data Provider and Consumer must answer the data subject's rights: • Information about the data processing (Article 12-14) • Access (Article 15) • Rectification (Article 16) • Erasure (Article 17) • Restriction on processing (Article 18) • Portability (Article 20) • Object (Article 21) • Not to be subject to a decision based solely on automated processing, including profiling (Article 22) • Lodge a complaint with the supervisory	Partially	This is under investigation to check how the HEDF can comply in future implementations.

	authority (Article 13).		
LEG-USR-024	Data Provider and Consumer must use the technical and organisational measures to protect personal data. Data provider and Consumer should use measures offered by TRUSTEE (Homomorphic encryption, SSI, authentication, federated learning, etc.).	Yes	These measures are taken within HEDF (homomorphic encryption, federated learning)
LEG-USR-027	All actors should be able to access only the data they actually need.	Yes	Filtering and suitable search criteria ensures correct data access.
LEG-USR-030	Data Provider may indicate to what countries the data can be made available	Yes	If such metadata are included, the HEDF will filter based on the Data Provider's choices.
LEG-USR-031	Data Provider may indicate that the data can be provided only to the Consumers from the countries with an adequate level of protection, as deemed by the European Commission.	Yes	If such metadata are included, the HEDF will filter based on the Data Provider's choices.
LEG-USR-056	Data Provider / Consumer / Model Provider being a public body should make dynamic data available through appropriate technical means immediately after their collection.	Partially	Future implementations of HEDF will make sure to retrieve only the data that users allow.
LEG-USR-058	Data Provider / Model Provider being a public body should make their research data as open as possible, as close as necessary, taking into account any concerns regarding, inter alia, IP rights, confidentiality, security and the protection of personal data. Research data shall be re-usable for commercial or non-commercial purposes in accordance with ODD, insofar as they are publicly funded and researchers, research performing organisations or research funding organisations have already made them publicly available through an institutional or subject-based repository.	Partially	Future implementations of HEDF will make sure to retrieve only the data that users allow.
LEG-USR-060	Consumer re-using the data from a public body is prohibited re-identifying any data subject to whom the data relates and must take technical and operational measures to prevent re-identification and to notify any data breach resulting in the re-identification of the data subjects concerned to the public sector body.	Yes	The HEDF will process any data in HE way, and only the data that stem from the agreement made for the usage of the dataset.
LEG-USR-066	Data Provider / Model Provider being a public sector body must, in accordance with	Yes	Data is always HE.

	law, ensure that the protected nature of data is preserved.		
LEG-USR-079	Data Provider / Model Provider / Developer being a "participant of data space offering data or data services to other participants" shall comply with the essential requirements to facilitate interoperability of data, data sharing mechanisms and services, defined in Article 28 Data Act: (a) the dataset content, use restrictions, licences, data collection methodology, data quality and uncertainty shall be sufficiently described in a machine-readable format to allow the recipient (Consumer) to find, access and use the data; (b) the data structures, data formats, vocabularies, classification schemes, taxonomies and code lists shall be described in a publicly available and consistent manner; (c) the technical means to access the data, such as application programming interfaces, and their terms of use and quality of service shall be sufficiently described to enable automatic access and transmission of data between parties, including continuously or in real-time in a machine-readable format where that is technically feasible and does not hamper the good functioning of the product; (d) the means to enable the interoperability of contracts for data sharing within their services and activities shall be provided.	Yes	The Semantic Representation layer uses open standards such as RDF semantic ontologies to facilitate interoperability and structured access to data
LEG-TSD-003	TRUSTEE must be developed as the system ensuring appropriate and secure processing of personal data. This is relevant especially for: a) the data considered as “special category data”, such as data concerning health or genetic data, as well as b) personal data concerning minors. These data will be processed by the future users of TRUSTEE (Data providers, Data consumers), especially in the HEALTH and EDUCATION domain.	Yes	No personal data will be processed.
LEG-TSD-004	With regards to the GDPR applicability: TRUSTEE could consider utilising anonymisation tools and provide appropriate explanation of applied anonymisation technique to the end users. However, it should be left up to the data providers	Yes	HE will be used for the processing of all data, so no raw data are accessed.

	whether the data which they would like to share via TRUSTEE should be anonymised or not – sometimes the anonymisation process causes loss of quality of the data set.		
LEG-TSD-005	TRUSTEE must be equipped with the technical and governance solutions which will allow for appropriate protection of the data which are or may be considered personal data, to ensure that the rights and freedoms of the data subjects (users or people whose data are accessed between users) are respected, can be exercised and will not be infringed.	Yes	HE offers data security and protection.
LEG-TSD-007	Security solutions recommended to be considered: - Backups, if possible encrypted and checked regularly should be made, to enable restoring the system. - Limitation of the size of the software stack used, - Installation of critical updates without delay, - Automation of a vulnerability watch, - Application of vulnerability detection tools, - Restriction of physical and software access to diagnostic and remote configuration ports, - Protection of the databases made available online, - Secure database management.	Yes	All these measures have been taken into consideration within HEDF.
LEG-TSD-012	With regard to the user data processed on TRUSTEE, TRUSTEE providers should implement solutions, such as technical and organisational measures (such as, in particular, encryption, pseudonymisation, privacy policy, and Terms of Use) which will address the requirements defined in the GDPR principles.	Yes	HE is used.
LEG-TSD-013	With regards to the TRUSTEE user's identification data, TRUSTEE providers will be controllers and below principles should be addressed by them: • Lawfulness: A legal basis in Article 6 (and where applicable an exception defined in Article 9 (2) GDPR for special categories of data) should be indicated. • Transparency: Information to the data subject (Data provider and Data consumer) about processing of their data by TRUSTEE	Yes	Data minimization techniques using HE have been addressed.

	Developers and maintaining parties should be presented. • Purpose limitation: Specific, explicit and legitimate purposes and process only data for this set purpose shall be defined. • Data Minimisation: Only data which is relevant and strictly necessary to provide the service can be processed. Pseudonymisation and/or anonymisation of data should be considered wherever possible. In the case of TRUSTEE, SSI authentication should allow for protection of users' privacy. Minimisation of the amount of data collected also in the log data and not storing sensitive or critical data must be applied. • Accuracy: Technical and organisational measures shall be put in place to ensure that the data provided by users is up to date and accurate and/or allow the users to correct inaccurate data. • Storage limitation: All data which is no longer necessary for the fulfilment of the purpose shall be deleted or anonymised, unless an exception applies (e.g., scientific research). Retention periods should be associated for each category of data (including logs). • Confidentiality, Integrity and Data Security: Technical and organisation measures, described in the sections below and in the technical requirements shall be implemented to mitigate risks and to ensure a high level of security, confidentiality and integrity of the data.		
LEG-TSD-014	With regards to the data provided onto TRUSTEE by data providers, which may be considered personal, the system should allow to address these principles as well, in particular: • Lawfulness: A legal basis in Article 6 (and where applicable Article 9 GDPR) should be indicated by the data provider as the basis for providing the data onto TRUSTEE and sharing them with the data consumers, and by the Data consumer for further processing for Data consumer's purposes. • Transparency: Data provider and Data consumer should be indicated as responsible	Yes	Data minimization techniques using HE have been addressed. The principle of confidentiality is addressed by ensuring that the data are homomorphically encrypted and system developers or system providers of TRUSTEE do not have access to raw data.

	for providing data subjects with the information about processing of their data on TRUSTEE and informed that the data consumers are among recipient of the data. • Purpose limitation: Data providers should indicate limitations which are in place in the context of the purposes for which the data provided onto TRUSTEE can be processed. • Data Minimisation: Only data which is relevant to achieve the defined purpose can be processed. Pseudonymisation and/or anonymisation of data should be considered wherever possible. In the case of TRUSTEE, homomorphic encryption is one of the means which should contribute to data providers' and data consumers' compliance with this requirement. • Accuracy: The end users should make sure that the technical and organisational measures shall be put in place on their sides as well as on TRUSTEE allow to keep the data accurate. The users providing the data should confirm whether and how the data provided onto TRUSTEE are accurate, by providing this information in the contract management layer and in the privacy impact assessment. • Storage limitation: All data which is no longer necessary for the fulfilment of the legitimate purpose set out by the Data providers or Data consumers shall be deleted or anonymised, unless an exception applies (e.g., scientific research). • Confidentiality, Integrity and Data Security: Technical and organisation measures, described in the sections below and in the technical requirements shall be implemented to mitigate risks to data processing and to ensure a high level of security, confidentiality and integrity of the data.		
LEG-TSD-021	The measures to be provided and which should be considered in the process of development and maintenance of TRUSTEE are: • Pseudonymisation and encryption of personal data – the TRUSTEE platform will employ homomorphic encryption technique;	Yes	Encryption is used for the data processing.

	• Measures to ensure the security of IT employed in the data processing, i.e. ensuring its ongoing confidentiality, integrity, availability and resilience – to answer these requirements, TRUSTEE will employ authentication services, such as SSI and SDK; • Incident response capabilities to restore the availability and access to personal data in a timely manner – due to the federated learning implementation, this responsibility would be rather dependent on the end-users providing datasets via TRUSTEE; • A regular assessment and evaluation of the effectiveness of the implemented technical and organisational security measures. The list is non-exhaustive. The TRUSTEE providers are encouraged to provide additional security measures to ensure appropriate level of data security.		
LEG-TSD-023	Logging systems should be used to detect anomalies or security events.	Yes	Logging is in place.
LEG-TSD-024	Penetration testing should be considered and performed.	Partially	Penetration testing will be conducted.
LEG-TSD-025	Efficient version control system should be set up, to track different versions of the code over the time, with appropriate authentication and permission system and regular backups, as well as efficient development system.	Yes	Github is used for documentation, backups and versioning.
LEG-TSD-026	Code quality metrics tools should be considered.	Partially	This has been taken into consideration.
LEG-TSD-027	Control of the use libraries and SDKs should be applied.	Partially	This has been taken into consideration.
LEG-TSD-028	When using free or open source software, TRUSTEE system developers should choose solutions that have an active community, good documentation and are regularly updated.	Yes	Python, Linux, PostgreSQL and other open source software is used.
LEG-TSD-029	When using types of solutions with commercial support, TRUSTEE system providers should contractually ensure maintenance and updates of the code during the whole lifecycle of the project.	Yes	So far, no commercial solutions are used, but if needed, HEDF will investigate this issue.
LEG-TSD-030	Change of the default configuration, audit, mapping of the dependencies of the SDKs and third party libraries should be considered.	Partially	This has been taken into consideration.

LEG-TSD-031	Implementation of cryptographic algorithms and protocols by TRUSTEE system developers by their own should be avoided, and choosing cryptographic libraries which are maintained, easy to use and recognised should be encouraged.	Partially	This has been taken into consideration.
LEG-TSD-032	Appropriate maintenance of libraries and SDKs should be considered, e.g. using dependency management system, managing updates to the dependencies, monitoring the versions of libraries and SDKs and regularly check the open source projects and their status, especially in the context of the change of domain or package ownership.	Partially	This has been taken into consideration.
LEG-TSD-033	Quality and security of the code and its documentation, as well as the documentation of the system architecture should be ensured.	Yes	This has been taken into consideration and Github is used for this purpose.
LEG-TSD-034	Applications must be tested (development test, security test). Real/ production data should not be used during the development and testing phase, as using personal data for testing purposes may go beyond their initial purpose.	Yes	Testings are carried out as well as last validations of the solutions.
LEG-TSD-035	When data security is comprised, TRUSTEE system providers must report the violation to the data controllers (Data providers and Data consumers) without undue delay if it is likely to pose a risk to the rights and freedoms of individuals.	Partially	This has been taken into consideration.

12.2. AIMAAS

Req id (WP2)	Req. Description	Coverage status (Yes/No/Partially)	How is addressed
LEG-USR-012	If other purpose is processing for scientific research based on the EU or national law, Data Provider must indicate that. In that case purpose is considered compatible by default, and appropriate technical and organizational	Yes	AIMaaS inherits TRUSTEE safeguards since it only operates on existing models or data descriptors within TRUSTEE.

	safeguards (measures) must be applied.		
LEG-USR-016	When a data subject exercises its right of erasure and/or restriction to the processing, Data Provider must ensure to remove the data from TRUSTEE / stop sharing the data of data subjects when requested by a data subject, required by the data protection laws and when there is no legal basis for further sharing and processing of the data.	Partially	This is under investigation, but is a requirement that will be fulfilled in the future implementations.
LEG-USR-018	Data Provider and Consumer must answer the data subject's rights: • Information about the data processing (Article 12-14) • Access (Article 15) • Rectification (Article 16) • Erasure (Article 17) • Restriction on processing (Article 18) • Portability (Article 20) • Object (Article 21) • Not to be subject to a decision based solely on automated processing, including profiling (Article 22) • Lodge a complaint with the supervisory authority (Article 13).	Partially	This is under investigation to check how the AIMaaS can comply in future implementations of its components and subcomponents. We inherit subject's right as all other TRUSTEE components
LEG-USR-024	Data Provider and Consumer must use the technical and organisational measures to protect personal data. Data provider and Consumer should use measures offered by TRUSTEE (Homomorphic encryption, SSI, authentication, federated learning, etc.).	Yes	These measures are taken within (SSI, HE, FL submodules and external ABBs
LEG-USR-030	Data Provider may indicate to what countries the data can be made available.	Yes	If such metadata are included, this module will filter based on the user's choices.

LEG-USR-031	Data Provider may indicate that the data can be provided only to the Consumers from the countries with an adequate level of protection, as deemed by the European Commission.	Yes	If such metadata are included, this module will filter based on the user's choices.
LEG-USR-068	If the Data Provider / Model Provider is going to be a "data holder" in the meaning of the Data Act (see legal framework), they should fulfil the obligations listed in Article 4 Data Act concerning, inter alia, way of making the data available to the user, the form in which the data should be made accessible, contractual agreement which may be concluded between user of a connected device/related service and Data provider, liability clauses, enabling identification through eID wallets, collecting minimum data about the user, protection of trade secrets (including construction of model contractual terms, confidential agreements, strict access protocols, technical standards), sharing of personal data and non-personal data does not apply to information derived or inferred from the data by complex algorithms combining output of multiple sensors.	Partially	This is under investigation to check how the AIMaaS can comply in future implementations of its components and subcomponents. We inherit subject's right as all other TRUSTEE components and we will completely support it in the next iteration
LEG-USR-074	If a Consumer (being "data recipient" in the meaning of the Data Act) has provided false information or deployed false information to the Data Provider ("data holder"), deployed deceptive or coercive means or abused evident gaps in the technical infrastructure of the Data Provider (data holder) designed to protect the data, has used the data made available for unauthorised purposes, including the	Partially	This is under investigation to check how the AIMaaS can comply in future implementations of its components and subcomponents. We inherit subject's right as all other TRUSTEE components and we will completely support it in the next iteration

	development of a competing product or has unlawfully disclosed data to another party, the Consumer (data recipient) shall be liable for the damages to the party suffering from the misuse or disclosure of such data and shall comply without undue delay with the requests of the Data Provider or the trade secret holder when they are not the same legal person to: (a) erase the data made available and any copies thereof; (b) end the production, offering, placing on the market or use of goods, derivative data or services produced on the basis of knowledge obtained through such data, or the importation, export or storage of infringing goods for those purposes, and destroy any infringing goods; (c) inform the user (data subject) of the unauthorised use or disclosure of the data and measures taken to put an end to the unauthorised use or disclosure of the data; (c) notify the Data Provider about the disclosure of such data.		
LEG-USR-081	Model Providers and Developers must determine if the AI provided by them is a high risk AI.	Yes	AIaaS/FL clearly complies with its implementation development technical characteristics. This is going to be highlighted in the UC
LEG-USR-082	If AI provided by the Model Provider / Developer is high risk, a risk management system must be established by them.	Partially	AIaaS/FL clearly complies with its implementation development technical characteristics. This is going to be highlighted in the UC
LEG-USR-083	If the system provided by Model Provider / Developer is high risk AI , the training, validation and	Partially	AIaaS/FL clearly complies with its implementation development technical characteristics. This is

	testing data sets must meet a certain quality criteria from Article 10, if the high-risk AI system uses techniques including the training of models with data. The data governance practices must concern in particular: relevant design choices, transparency as regards the original purpose of data collection, data collection process, data preparation operations (annotation, labelling, cleaning, updating, enrichment and aggregation), the formulation of assumptions, an assessment of the availability, quantity and suitability of the needed datasets, examination in view of possible biases, appropriate measures to detect, prevent and mitigate possible biases, identification of data gaps or shortcomings and how they can be addressed. At the level of individual datasets and their combination, the training, validation and testing datasets (incl. labels) shall be relevant, representative, appropriately vetted for errors and complete, have appropriate statistical properties, including, as regards the persons or groups of persons in relation to whom the high-risk AI system is intended to be used.		going to be highlighted in the UC
LEG-USR-084	If AI provided by the Model Provider / Developer is high risk, the events in the AI system must be automatically recorded (logs must be kept), to enable tracing of AI system's functionality and to monitor the situations posing risks.	Partially	AIaaS/FL clearly complies with its implementation development technical characteristics. This is going to be highlighted in the UC
LEG-USR-085	If AI provided by the Model Provider /Developer is high risk, a technical documentation of the high-risk AI system must be	Partially	AIaaS/FL clearly complies with its implementation development technical characteristics. This is

	provided before it is placed on the market or put into service, and regularly updated.		going to be highlighted in the UC
LEG-USR-086	If AI provided by the Model Provider / Developer is high risk, Model Provider / Developer must ensure that their (AI systems') operation is sufficiently transparent to enable users (Consumers) to interpret the system's output and use it appropriately, by providing instructions in an appropriate digital format and information about the purpose, capabilities and limitations of the AI system in a clear manner. Specific parts of the information are included in Article 13 AIA. Additionally, the AI systems must be designed and developed in a way that they can be effectively overseen by natural persons during the time the AI system is in use. This refers also to appropriate human-machine interface tools.	Partially	AIaaS/FL clearly complies with its implementation development technical characteristics. This is going to be highlighted in the UC
LEG-USR-087	If AI provided by the Model Provider / Developer is high risk, Model Provider / Developer shall ensure appropriate level of accuracy, robustness, cybersecurity. The levels of accuracy and accuracy metrics shall be declared in the instructions of use. The AI systems must be adaptable against errors, faults and inconsistencies that may occur within the system or the environment in which it operates and withstand any attempts by unauthorised parties to alter the use or performance of the AI system.	Partially	AIaaS/FL clearly complies with its implementation development technical characteristics. This is going to be highlighted in the UC
LEG-USR-088	If AI provided by the Model Provider / Developer is high risk,	Partially	AIaaS/FL clearly complies with its implementation development

	Model Provider / Developer must establish and document a post-monitoring plan for high-risk AI systems to actively and systematically collect, document and analyse data provided by users or other data collected on the performance of the AI system throughout their lifecycle.		technical characteristics. This is going to be highlighted in the UC
LEG-USR-089	Model Provider / Developer must provide appropriate quality management system and implement a conformity assessment procedure for the AI system before it is placed on the market.	Partially	AIaaS/FL clearly complies with its implementation development technical characteristics. This is going to be highlighted in the UC
LEG-USR-090	Model Provider / Developer who provides AI system (not only high-risk AI system) must communicate clearly to a user being a natural person that this person interacts with an AI system, unless evident from the context.	Partially	AIaaS/FL clearly complies with its implementation development technical characteristics. This is going to be highlighted in the UC
LEG-USR-091	Model Provider / Developer and Consumer using models must adhere to the <u>requirements for Trustworthy AI: Human Agency and Oversight, Technical Robustness and Safety, Privacy and Data Governance, Transparency, Diversity, Non-discrimination and Fairness, Environmental and Societal Well-Being, Accountability, further described in the legal framework and in socio-ethical requirements section.</u>	Partially	AIaaS/FL clearly complies with its implementation development technical characteristics. This is going to be highlighted in the UC
LEG-TSD-001	Should the TRUSTEE system developers and parties maintaining the system at any point handle or process any type of personal data, they should balance their legitimate interest with the	Yes	If any personal data is stored to this module, will only be accessed by the authorised parties and provide them the

	fundamental rights and freedoms of the data subjects whose data they will be processing.		ability to delete the data when required.
LEG-TSD-002	TRUSTEE must have the fundamental rights principles at its core, by ensuring that the privacy of the individuals is respected and protected.	Yes	This module operated within TRUSTEE secure “trusted” environment and only through the included privacy mechanisms
LEG-TSD-006	Any website, application or server must incorporate basic state-of-the-art security rules, not only on network communications but also on authentication and infrastructure.	Yes	AIaaS fully supports secure web services and application as agreed within TRUSTEE
LEG-TSD-007	Security solutions recommended to be considered: - Backups, if possible encrypted and checked regularly should be made, to enable restoring the system. - Limitation of the size of the software stack used, - Installation of critical updates without delay, - Automation of a vulnerability watch, - Application of vulnerability detection tools, - Restriction of physical and software access to diagnostic and remote configuration ports, - Protection of the databases made available online, - Secure database management.	Yes	AIaaS fully supports secure web services and application as agreed within TRUSTEE
LEG-TSD-019	TRUSTEE must incorporate the solutions compliant with data protection requirements already in planning and development phases (privacy by design), and ensure that only personal data necessary	Yes	TRUSTEE AIMaaS will not expose data to unauthorized users.

	for the purpose of providing TRUSTEE to the users are processed (privacy by default).		
LEG-TSD-023	Logging systems should be used to detect anomalies or security events.	Partially	Loggin is partially implemented
LEG-TSD-024	Penetration testing should be considered and performed.	Partially	Penetration testing will be conducted.
LEG-TSD-025	Efficient version control system should be set up, to track different versions of the code over the time, with appropriate authentication and permission system and regular backups, as well as efficient development system.	Yes	Github is used for documentation, backups and versioning.
LEG-TSD-026	Code quality metrics tools should be considered.	Partially	We conduct unit quality code control within the development teams
LEG-TSD-027	Control of the use libraries and SDKs should be applied.	Yes	We have full control of the use of SDKs and libraries
LEG-TSD-028	When using free or open source software, TRUSTEE system developers should choose solutions that have an active community, good documentation and are regularly updated.	Partially	At the best of our abilities we use open source software
LEG-TSD-029	When using types of solutions with commercial support, TRUSTEE system providers should contractually ensure maintenance and updates of the code during the whole lifecycle of the project.	Partially	So far, no commercial solutions are used, but if needed we will investigate this issue.
LEG-TSD-030	Change of the default configurations, audit, mapping of the dependencies of the SDKs and	Yes	This is performed on a need to be basis and is logged

	third party libraries should be considered.		
LEG-TSD-032	Appropriate maintenance of libraries and SDKs should be considered, e.g. using dependency management system, managing updates to the dependencies, monitoring the versions of libraries and SDKs and regularly check the open source projects and their status, especially in the context of the change of domain or package ownership.	Yes	We follow the common consensus of the consortium and we keep our own maintained internal processes
LEG-TSD-033	Quality and security of the code and its documentation, as well as the documentation of the system architecture should be ensured.	Yes	This has been taken into consideration and Github is used for this purpose.
LEG-TSD-034	Applications must be tested (development test, security test). Real/ production data should not be used during the development and testing phase, as using personal data for testin purposes may go beyond their initial purpose.	Yes	Testings are carried out as well last validations of the solutions.
LEG-TSD-035	When data security is comprised, TRUSTEE system providers must report the violation to the data controllers (Data providers and Data consumers) without undue delay if it is likely to pose a risk to the rights and freedoms of individuals.	Partially	This is not fully implemented yet we will deliver it in the next iterations.
LEG-TSD-041	Explore whether any specific addressing of tools/considerations of allowing for the sharing of data/documents held by public sector bodies could be facilitated by TRUSTEE.	Partially	Tools can be explored (if applicable).

LEG-TSD-092	TRUSTEE system developers should determine whether, by allowing sharing the AI solutions by the Model providers (users providing AI technology to Data providers and Data consumers), they also participate in providing AI systems. If so, AI Act requirements will apply in the future.	Yes	All systems/subsystems are built accordingly
LEG-TSD-093	If TRUSTEE system developers are considered AI system providers , TRUSTEE system developers must determine if the AI is a high risk AI .	Partially	All systems/subsystems are built accordingly
LEG-TSD-094	If TRUSTEE system developers occur to be providers of high risk AI , a risk management system must be established, implemented, documented and maintained. Further details are included in the AIA.	Partially	All systems/subsystems are built accordingly
LEG-TSD-095	If TRUSTEE system developers occur to be providers of high risk AI , the training, validation and testing data sets must meet a certain quality criteria from Article 10, if the high-risk AI system uses techniques including the training of models with data. The data governance practices must concern in particular: relevant design choices, transparency as regards the original purpose of data collection, data collection process, data preparation operations (annotation, labelling, cleaning, updating, enrichment and aggregation), the formulation of assumptions, an assessment of the availability, quantity and suitability of the needed datasets, examination in	Partially	All systems/subsystems are built accordingly

	view of possible biases, appropriate measures to detect, prevent and mitigate possible biases, identification of data gaps or shortcomings and how they can be addressed. At the level of individual datasets and their combination, the training, validation and testing datasets (incl. labels) shall be relevant, representative, appropriately vetted for errors and complete, have appropriate statistical properties, including, as regards the persons or groups of persons in relation to whom the high-risk AI system is intended to be used.		
LEG-TSD-096	If TRUSTEE system developers occur to be providers of high risk AI, the events in the AI system must be automatically recorded (logs must be kept), to enable tracing of AI system's functionality and to monitor the situations posing risks.	Partially	All systems/subsystems are built accordingly
LEG-TSD-097	If TRUSTEE system developers occur to be providers of high risk AI, a technical documentation of the high-risk AI system must be provided before it is placed on the market or put into service, and regularly updated.	Partially	All systems/subsystems are built accordingly
LEG-TSD-098	If TRUSTEE system developers occur to be providers of high risk AI, TRUSTEE system developers providing AI systems must ensure that their (AI systems') operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately, by providing instructions in an appropriate digital format and information about the purpose,	Partially	All systems/subsystems are built accordingly

	capabilities and limitations of the AI system in a clear manner. Specific parts of the information are included in Article 13 AIA. Additionally, the AI systems must be designed and developed in a way that they can be effectively overseen by natural persons during the time the AI system is in use. This refers also to appropriate human-machine interface tools.		
LEG-TSD-099	If TRUSTEE system developers occur to be providers of high risk AI, shall ensure appropriate level of accuracy, robustness, cybersecurity. The levels of accuracy and accuracy metrics shall be declared in the instructions of use. The AI systems must be adaptable against errors, faults and inconsistencies that may occur within the system or the environment in which it operates and withstand any attempts by unauthorised parties to alter the use or performance of the AI system.	Partially	All systems/subsystems are built accordingly
LEG-TSD-100	If TRUSTEE system developers occur to be providers of high risk AI, TRUSTEE system developers must establish and document a post-monitoring plan for high-risk AI systems to actively and systematically collect, document and analyse data provided by users or other data collected on the performance of the AI system throughout their lifecycle	Partially	All systems/subsystems are built accordingly
LEG-TSD-101	If TRUSTEE system developers occur to be providers of high risk AI, TRUSTEE system developers must provide appropriate quality management system and implement a conformity assessment procedure	Partially	All systems/subsystems are built accordingly

	for the AI system before it is placed on the market.		
LEG-TSD-102	If TRUSTEE system developers occur to be providers of high risk AI , TRUSTEE system developers must communicate clearly that a user being a natural person interacts with an AI system, unless evident from the context.	Partially	All systems/subsystems are built accordingly
LEG-TSD-103	If TRUSTEE system developers occur to be providers of high risk AI , TRUSTEE system developers must adhere to the requirements for Trustworthy AI: Human Agency and Oversight, Technical Robustness and Safety, Privacy and Data Governance, Transparency, Diversity, Non-discrimination and Fairness, Environmental and Societal Well-Being, Accountability, further described in the legal framework and in socio-ethical requirements section.	Partially	All systems/subsystems are built accordingly
LEG-TSD-107	TRUSTEE system developers must abstain from designing and operating their interfaces in a manipulative or deceitful manner or in a way hindering service recipients' ability to make independent and informed choices. (This is an additional requirement for providers of online platforms, under the category of hosting services. To the extent that TRUSTEE's activities and service provision falls under this category, this could be applied.)	Partially	All systems/subsystems are built accordingly
LEG-TSD-108	TRUSTEE should aim to transpose transparency obligations for contracts between Data providers/Data consumers being business entities and TRUSTEE	Partially	All systems/subsystems are built accordingly

	service providers, as identified in the legal framework.		

12.3. TAI-SDF

TRUSTEE's user legal requirements were listed in the deliverable D2.1. The compliance of TAI-SDF with the applicable legal requirements for system developers and end-users is presented in Table 28: TAI-SDF – Legal requirements for system developers and Table 29: TAI-SDF – Legal requirements for end-users.

Table 28: TAI-SDF – Legal requirements for system developers

Req id (WP2)	Req. Description	Compliance (Yes/No/Partially)	How is addressed
LEG-TSD-006	Any website, application or server must incorporate basic state-of-the-art security rules, not only on network communications but also on authentication and infrastructure.	Yes	Requirement is applicable and already reflected into technical requirements.
LEG-TSD-013	The outcome of the compatibility assessment should be indicated by Data Provider, preferably in a data sharing agreement concluded between Data provider and Data consumer.	Partially	Requirement is partially applicable in cooperation with STM. This will be further investigated, and reply will be provided in D3.2.
LEG-TSD-016	When a data subject exercises its right of erasure and/or restriction to the processing, Data Provider must ensure to remove the data from TRUSTEE / stop sharing the data of data subjects when requested by a data subject, required by the data protection laws and when there is no legal basis for further sharing and processing of the data.	Partially	This will be further investigated, and reply provided in D3.2.

LEG-TSD-019	TRUSTEE must incorporate the solutions compliant with data protection requirements already in planning and development phases (privacy by design), and ensure that only personal data necessary for the purpose of providing TRUSTEE to the users are processed (privacy by default).	Partially	Since TAI-SDF will process userID and datasetID, a question related to management of personal data will be added in TAI-SDF questionnaire. This will be further investigated in the scope of D3.2.

Table 29: TAI-SDF – Legal requirements for end-users

Req id (WP2)	Req. Description	Compliance (Yes/No/Partially)	How is addressed
LEG-USR-019	Data provider and Data consumer must indicate whether application of AI models offered by TRUSTEE will lead to any automated individual decision-making about a data subject whose data are subject to processing.	Partially	This requirement could be translated into a question of a checklist of module “Security & privacy”. At the moment the question has been included in GDPR checklist (Annex 5).
LEG-USR-020	Data provider and Data consumer must indicate if data subject (whose homomorphically encrypted data is accessed and processed) has been informed about automated individual decision making.	Partially	This requirement could be translated into a question of a checklist of module “Security & privacy”. At the moment the question is included in GDPR checklist (Annex 5).
LEG-USR-021	Data provider and Data consumer must provide what is the exception applicable, which allows the automated individual decision making (necessity to perform a contract, authorisation by the EU or national law, explicit consent of a data subject).	Partially	This requirement could be translated into a question of a checklist of module “Security & privacy”. At the moment the question is included in GDPR checklist (Annex 5).
LEG-USR-022	For special categories of data, Data provider and Data consumer must indicate whether automated	Partially	This requirement could be translated into a question of a checklist of module “Security &

	individual decision making is allowed due to an explicit consent of a data subject or due to a substantial public interest, based on the EU or national law.		privacy”. At the moment the question is included in GDPR checklist (Annex 5).
LEG-USR-023	Data provider and Data consumer applying AI models must perform a DPIA.	Yes	This requirement could be translated into a question of a checklist of module “Compliance”. At the moment the question is included in GDPR checklist (Annex 5).
LEG-USR-080	Data provider / Model provider / Developer offering smart contract to Data consumer in the context of an agreement to make data available shall comply with the essential requirements defined in Article 30 Data Act, concerning robustness and access control, safe termination and interruption, equivalence, protection of confidentiality of trade secrets, but also auditability.	Yes	TAI-SDF supports the assessment of AI robustness by providing a dedicated questionnaire on the robustness. Other aspects requested by this requirement will be further investigated in the scope of D3.2.
LEG-USR-081	Model providers and Developers must determine if the AI provided by them is a high risk AI.	Partially	TAI-SDF does not have any risk analysis functionalities at the moment. To investigate if adding a dedicated checklist is feasible. At the moment general questions pertaining to the risk assessment are included in the questionnaire in Annex 4.
LEG-USR-082	If AI provided by the Model provider / Developer is high risk, a risk management system must be established by them.	Partially	TAI-SDF does not have any risk analysis functionalities at the moment. To investigate if adding a dedicated checklist is feasible. As of now, general questions pertaining to the risk assessment are included in the questionnaire in Annex 4.

LEG-USR-083	If the system provided by Model provider / Developer is high risk AI, the training, validation and testing data sets must meet a certain quality criteria from Article 10, if the high-risk AI system uses techniques including the training of models with data. The data governance practices must concern in particular: relevant design choices, transparency as regards the original purpose of data collection, data collection process, data preparation operations (annotation, labelling, cleaning, updating, enrichment and aggregation), the formulation of assumptions, an assessment of the availability, quantity and suitability of the needed datasets, examination in view of possible biases, appropriate measures to detect, prevent and mitigate possible biases, identification of data gaps or shortcomings and how they can be addressed. At the level of individual datasets and their combination, the training, validation and testing datasets (incl. labels) shall be relevant, representative, appropriately vetted for errors and complete, have appropriate statistical properties, including, as regards the persons or groups of persons in relation to whom the high-risk AI system is intended to be used.	Yes	TAI-SDF questionnaire, assessing the trustworthiness of the AI system (see Annex 4), addresses these aspects.
LEG-USR-084	If AI provided by the Model provider / Developer is high risk, the events in the AI system must be automatically recorded (logs must be kept), to enable tracing of AI system's functionality and to monitor the situations posing risks.	Partially	TAI-SDF questionnaire, assessing the trustworthiness of the AI system (see Annex 4), addresses these aspects. Please note that the compliance for this requirement is Partially since this requirement need to be addressed by AIMaaS

LEG-USR-085	If AI provided by the Model provider /Developer is high risk, a technical documentation of the high-risk AI system must be provided before it is placed on the market or put into service, and regularly updated.	Partially	TAI-SDF questionnaire, assessing the trustworthiness of the AI system (see Annex 4), addresses these aspects. Please note that the compliance for this requirement is Partially since this requirement need to be addressed by AIMaaS.
LEG-USR-086	If AI provided by the Model provider / Developer is high risk, Model provider / Developer must ensure that their (AI systems') operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately, by providing instructions in an appropriate digital format and information about the purpose, capabilities and limitations of the AI system in a clear manner. Specific parts of the information are included in Article 13 AIA. Additionally, the AI systems must be designed and developed in a way that they can be effectively overseen by natural persons during the time the AI system is in use. This refers also to appropriate human-machine interface tools.	Yes	TAI-SDF questionnaire, assessing the trustworthiness of the AI system (see Annex 4), addresses these aspects.
LEG-USR-087	If AI provided by the Model provider / Developer is high risk, Model provider / Developer shall ensure appropriate level of accuracy, robustness, cybersecurity. The levels of accuracy and accuracy metrics shall be declared in the instructions of use. The AI systems must be adaptable against errors, faults and inconsistencies that may occur within the system or the environment in which it operates and withstand any attempts by	Partially	TAI-SDF questionnaire, assessing the trustworthiness of the AI system (see Annex 4), addresses these aspects. Please note that the compliance for this requirement is Partially since this requirement need to be addressed by AIMaaS.

	unauthorised parties to alter the use or performance of the AI system.		
LEG-USR-088	If AI provided by the Model provider / Developer is high risk, Model provider / Developer must establish and document a post-monitoring plan for high-risk AI systems to actively and systematically collect, document and analyse data provided by users or other data collected on the performance of the AI system throughout their lifecycle.	Partially	TAI-SDF questionnaire, assessing the trustworthiness of the AI system (see Annex 4), addresses these aspects. Please note that the compliance for this requirement is Partially since this requirement need to be addressed by AIMaaS.
LEG-USR-089	Model provider / Developer must provide appropriate quality management system and implement a conformity assessment procedure for the AI system before it is placed on the market.	Yes	TAI-SDF supports the quality and conformity assessment by design. TAI-SDF questionnaire, assessing the trustworthiness of the AI system (see Annex 4), addresses these aspects.
LEG-USR-090	Model provider / Developer who provides AI system (not only high-risk AI system) must communicate clearly to a user being a natural person that this person interacts with an AI system, unless evident from the context.	Yes	TAI-SDF questionnaire, assessing the trustworthiness of the AI system (see Annex 4), addresses these aspects.
LEG-USR-091	Model provider / Developer and Data consumer using models must adhere to the requirements for Trustworthy AI: Human Agency and Oversight, Technical Robustness and Safety, Privacy and Data Governance, Transparency, Diversity, Non-discrimination and Fairness, Environmental and Societal Well-Being, Accountability, further described in the legal framework and in socio-ethical requirements section.	Yes	TAI-SDF is built upon Trustworthiness aspects. Questionnaire presented in Annex 4 has been built on Trustworthy AI Assessment List.
LEG-USR-092	Model provider/ Developer / Data consumer using the AI model	Partially	TAI-SDF questionnaire, assessing the trustworthiness of

	should consider conducting fundamental rights impact assessment (in addition to a DPIA) as part of AI impact assessment.		the AI system (see Annex 4), addresses relevant aspects concerning fundamental rights of the individuals. It will be further investigated if and how to extend it.
LEG-USR-093	Model provider and Developer, when developing, and Data consumer, when using algorithms for the use in HEALTH domain, must consider, in addition to the requirements above, fulfilling the following: -preserving autonomy of a data subject, - ensure safety, accuracy and efficiency, quality control and improvement (impacting human well-being, human safety and the public interest); - ensure that the system is intelligible and understandable to developers, users and regulators (transparency, explainability and intelligibility); - consider responsibility for the AI system and ensure that AI relies on certified sources and intangible data with models conceived in a multi-disciplinary manner, in a secure technological environment (responsibility and accountability); - consider different languages, ability and forms of communications, to make AI system as inclusive as possible and non-discriminatory, elimination of unintended bias in the design, implementation and use of the algorithms (inclusiveness and equity); - observe how developed AI answers to communicated expectations and requirements in the context in which it is used.	Yes	TAI-SDF questionnaire, assessing the trustworthiness of the AI system (see Annex 4), addresses these aspects.

12.4. ATR

Req id (WP2)	Req. Description	Coverage status (Yes/No/Partially)	How is addressed
LEG-TSD-001	Should the TRUSTEE system developers and parties maintaining the system at any point handle or process any type of personal data, they would be expected to balance their legitimate interests with the fundamental rights and freedoms of the data subjects whose data they will be processing.	Yes	If any personal data is stored to ATR, it will only be accessed by the authorised parties and provide them the ability to delete the data when required.
LEG-TSD-002	TRUSTEE system must have the fundamental rights principles at its core, by ensuring that the privacy of the individuals is respected and protected.	Ye	Permissioned blockchain ensures full privacy.
LEG-TSD-003	TRUSTEE must be developed as the system ensuring appropriate and secure processing of personal data. This is relevant especially for: a) the data considered as “special category data”, such as data concerning health or genetic data, as well as b) personal data concerning minors. These data will be processed by the future users of TRUSTEE (Data providers, Data consumers), especially in the HEALTH and EDUCATION domain	Yes	Requirement can be applicable if required to store sensitive data utilising the “Private data collection” feature of the ATR
LEG-TSD-006	Any website, application or server must incorporate basic state-of-the-art security rules, not only on network communications but also	Yes	The security approaches, such as permissioned network, robust access control, off-chain data storage, etc., are in place ensuring

	on authentication and infrastructure.		only authorised entities can communicate.
LEG-TSD-007	Security solutions recommended to be considered: - Backups, if possible encrypted and checked regularly should be made, to enable restoring the system. - Limitation of the size of the software stack used, - Installation of critical updates without delay, - Automation of a vulnerability watch, - Application of vulnerability detection tools, - Restriction of physical and software access to diagnostic and remote configuration ports, - Protection of the databases made available online, - Secure database management.	Yes	All types of security solutions, such as encryption, endorsement policies, vulnerability detection tools, etc., will be in place and considered to ensure the integrity of the recorded transactions.
LEG-TSD-012	With regard to the user data processed on TRUSTEE, TRUSTEE providers should implement solutions, such as technical and organisational measures (such as in particular, encryption, pseudonymisation, privacy policy, and terms of use) which will address the requirements defined in the GDPR principles.	Yes	The privacy preserving policy/feature of ATR will establish an extra layer of security and will not only record transactions to the immutable storage but also off-chain storage giving the ability for data to be changed or deleted and thus fully compliant with GDPR principles.
LEG-TSD-014	With regards to the TRUSTEE user's identification data, TRUSTEE system providers will be controllers and below	Partially	Confidentiality, Integrity and data security will be maintained. Consensus mechanism, channel-based privacy, private data collections, Public Key

	principles should be addressed by them: • Lawfulness: A legal basis in Article 6 (and where applicable an exception defined in Article 9 (2) GDPR for special categories of data) should be indicated. • Transparency: Information to the data subject (Data provider and Consumer) about processing of their data by TRUSTEE Developers and maintaining parties should be presented. • Purpose limitation: Specific, explicit and legitimate purposes and process only data for this set purpose shall be defined. • Data minimisation: Only data which is relevant and strictly necessary to provide the service can be processed. Pseudonymisation and/or anonymisation of data should be considered wherever possible. In the case of TRUSTEE, SSI authentication should allow for protection of users' privacy. Minimisation of the amount of data collected also in the log data and not storing sensitive or critical data must be applied. • Accuracy: Technical and organisational measures shall be put in place to ensure that the data provided by users is up to date and accurate and/or allow the users to correct inaccurate data. • Storage limitation: All data which is no longer necessary for the fulfilment of the purpose shall be deleted or anonymised, unless an exception applies (e.g., scientific research). Retention periods should be associated for		Infrastructure (PKI) and digital certificated, data encryption through cryptographic protocols, etc will ensure the above. The unnecessary data can be deleted from the off-chain data storage and also shared on a need-to-know basis. Furthermore, the endorsement policies can define which particular blockchain organisations can take part in the endorsement, thus eliminating data exposure to unnecessary entities.

	each category of data (including logs). Confidentiality, Integrity and Data Security: Technical and organisation measures, described in the sections below and in the technical requirements shall be implemented to mitigate risks and to ensure a high level of security, confidentiality and integrity of the data.		
LEG-TSD-021	The measures to be provided and which should be considered in the process of development and maintenance of TRUSTEE are: - Pseudonymisation and encryption of personal data – the TRUSTEE platform will employ homomorphic encryption technique; - Measures to ensure the security of IT employed in the data processing, i.e. ensuring its ongoing confidentiality, integrity, availability and resilience – to answer these requirements, TRUSTEE will employ authentication services, such as SSI and SDK; - Incident response capabilities to restore the availability and access to personal data in a timely manner – due to the federated learning implementation, this responsibility would be rather dependent on the end-users providing datasets via TRUSTEE; - A regular assessment and evaluation of the effectiveness of the implemented technical and organisational security measures. The list is non-exhaustive. The TRUSTEE providers are	Yes	Personal data will be stored securely at off-chain storage through the “Private Data Collection” feature and will have access to only authorised ABB’s. Thus, full confidentiality will be maintained. Also, integration with external security tools, e.g., Capiler, with Hyperledger Fabric will allow to conduct regular assessment of the deployed decentralised network.

	encouraged to provide additional security measures to ensure appropriate level of data security.		
LEG-TSD-023	Logging systems should be used to detect anomalies or security events.	Yes	One of the features of Hyperledger Fabric is to keep the log.
LEG-TSD-024	Penetration testing should be considered and performed.	Partially	Various benchmarking tools, such as Hyperledger Caliper, will be used to determine the effectiveness. Besides that, network scanning tools can be used to determine services in the network, for instance, open ports and services running on nodes. Additionally, the smart contracts can be assessed for vulnerabilities such as tampering with transactions, etc.
LEG-TSD-025	Efficient version control system should be set up, to track different versions of the code over the time, with appropriate authentication and permission system and regular backups, as well as efficient development system.	Yes	All the chaincodes can be backed up and versioned utilising version control system. Additionally, Hyperledger Fabric comprises this basic functionality which allows installing updated versions of Chaincode.
LEG-TSD-026	Code quality metrics tools should be considered.	Yes	Hyperledger Fabric provides guidelines and best practices for chaincode. "Hyperledger Fabric Documentation" can be followed to maintain the quality.
LEG-TSD-027	Control of the use libraries and SDKs must be applied.	Yes	The Fabric SDK is provided by the Hyperledger project. It is hosted by the Linux Foundation and offers a set of libraries and tools which can be leveraged to develop applications that can

			interact with the Hyperledger Fabric permissioned blockchain network to record transactions.
LEG-TSD-028	When using free or open source software, TRUSTEE system developers should choose solutions that have an active community, good documentation and are regularly updated.	Yes	Hyperledger Fabric provides guidelines and documentation which are regularly updated. The documentation will be followed.
LEG-TSD-029	When using types of solutions with commercial support, TRUSTEE system providers should contractually ensure maintenance and updates of the code during the whole lifecycle of the project.	Yes	Hyperledger Fabric allows to re-install updated versions of chaincode, and also chaincodes with different functionalities in multiple peers. Thus, it provides full flexibility with the maintenance and update of the code.
LEG-TSD-030	Change of the default configurations, audit, mapping of the dependencies of the SDKs and third party libraries should be considered.	Yes	Third party libraries or frameworks such as CouchDB, gRPC, OpenSSL, etc. can be leveraged by Hyperledger Fabric to provide specific functionalities.
LEG-TSD-031	Implementation of cryptographic algorithms and protocols by TRUSTEE system developers by their own should be avoided, and choosing cryptographic libraries which are maintained, easy to use and recognised should be encouraged.	Yes	Hyperledger Fabric is incorporated with well recognised cryptographic libraries and components, such as SHA-256, AES, ECDSA, OpenSSL, BSSSP, etc., ensuring enhanced security, trust and reliability.
LEG-TSD-032	Appropriate maintenance of libraries and SDKs should be considered, e.g. using dependency management system, managing updates to the dependencies, monitoring the versions of libraries and SDKs and regularly check the open source projects	Yes	The Hyperledger Fabric official documentation is updated regularly providing guidelines on the maintenance details and on the versions of libraries and other information. The maintenance can

	and their status, especially in the context of the change of domain or package ownership.		be carried out as per the documentation.
LEG-TSD-033	Quality and security of the code and its documentation, as well as the documentation of the system architecture must be ensured.	Yes	Hyperledger Fabric provides guidelines and best practices for chaincode, libraries, etc. The guideline will be followed to maintain the quality and security.
LEG-TSD-034	Applications must be tested (development test, security test). Real/ production data should not be used during the development and testing phase, as using personal data for testing purposes may go beyond their initial purpose.	Yes	The chaincode will go through rigorous testing process according to the guidelines in order to meet the production requirements. Additionally, it will ensure that the deletion of personal data is fully functional and implemented through chaincode.
LEG-TSD-035	When data security is compromised, TRUSTEE system providers must report the violation to the data controllers (Data Providers and Data Consumers) without undue delay if it is likely to pose a risk to the rights and freedoms of individuals.	Yes	The integrated tools, e.g., Hyperledger Explorer, etc., will trigger notifications and provide live updates of the inner functioning enabling to gain insight of all the network components. It is the Dashboard which will be developed specifically for monitoring purposes will comprise the functionalities that will allow to detect and report anomalies.
LEG-TSD-036	In case of the transfer of the data via TRUSTEE to third (non-EU) countries, appropriate safeguards must be considered and TRUSTEE should enable their implementation by the user.	Partially	SHA-256 will be used to hash the transactional data ensuring integrity and uniqueness. Additionally, a robust access control will be in place enabling only the authorised parties, regardless of their location, to access the recorded transactions.
LEG-TSD-041	Explore whether any specific addressing of tools/considerations of allowing for the sharing of data/documents held by public	Partially	Tools can be explored (if applicable). It is to be further explored in next phases of the TRUSTEE activities.

	sector bodies could be facilitated by TRUSTEE.		
LEG-TSD-042	TRUSTEE system developers should identify and take appropriate and proportionate technical and organizational measures to manage risks created by the security systems.	Yes	Hyperledger Fabric comprises with built-in security functionalities that help managing risks.
LEG-TSD-043	The level of security which should be ascertained has to be appropriate to the risks posed, and should take into account the following elements: a. the security of systems and facilities; b. incident handling; c. business continuity management; d. monitoring, auditing and testing; and d. compliance with international standards.	Yes	Hyperledger Fabric comprises with built-in security functionalities that help managing risks. Additionally, the integrated tools will provide monitoring, auditing and other security functionalities.
LEG-TSD-044	The security measures taken by TRUSTEE system developers shall include: (a) detection processes and procedures maintained and tested to ensure timely and adequate awareness of anomalous events; (b) processes and policies on reporting incidents and identifying weaknesses and vulnerabilities in their information systems; (c) a response in accordance with established procedures and reporting the results of the measure taken; (d) an assessment of the incident's severity, documenting knowledge from incident analysis and collection of	Yes	All security measures will be adopted as per the guidelines.

	relevant information which may serve as evidence and support a continuous improvement process.		
LEG-TSD-045	TRUSTEE system developers must fulfil the obligation of monitoring, auditing and testing, there should be policies implemented which will define supervision, inspection and verification procedures.	Yes	Hyperledger Fabric comprises with built-in security functionalities, such as identity management, access control, ledger encryption, consensus mechanism, channel isolation, endorsement policies, etc., help managing risks. Additionally, the integrated Hyperledger Explorer will provide monitoring, auditing and other functionalities. Also, Hyperledger Fabric provides guidelines/documentation and best practices for chaincode, libraries, etc. The guidelines can be followed to maintain the required obligation.
LEG-TSD-046	TRUSTEE providers may be recognised either as essential (as provider of trusted services) or as important (as research institutions) entities, therefore they should adhere, in all phases of the development of TRUSTEE and its further maintenance, to the NIS 2 Directive stemming obligations, which will be further transposed and established at the national level.	Yes	The documentation provided by the Hyperledger Fabric can be followed to gain insight on the policies, requirements, etc.
LEG-TSD-047	TRUSTEE system developers should consider and implement the requirements stemming from the NIS and NIS 2 Directive, to ensure appropriate level of security of information, especially implementation of the means such as:	Yes	All security measures can be adopted as per the Hyperledger Fabric documentation. The documentation provides information and discusses the implementation of various methods that can enhance security.

	- policies on risk analysis and information system security, - incident handling, - business continuity , - supply chain security, - security in acquisition, development and maintenance of network and information system, - assessment of the effectiveness of implemented measures, - cybersecurity training and basic cyber hygiene practices, - procedures regarding the use of cryptography and, where appropriate, encryption, access control policies, asset management and HR security, - the use of multi-factor authentication or continuous authentication solutions within the entity, where appropriate.		
LEG-TSD-048	TRUSTEE system developers shall implement appropriate incident reporting procedures.	Yes	Integration of security tools will provide reports on security analysis.
LEG-TSD-050	TRUSTEE providers shall thoroughly plan and implement appropriate security measures and constantly monitor, test and audit them, as well as keep the documentation describing those measures in a transparent and clear manner.	Yes	The security model of Hyperledger covers these aspects and can be followed to plan and implement appropriate security measures.
LEG-TSD-076	If TRUSTEE happens to be intermediation service according to the DGA, TRUSTEE service providers must inform data holders (data providers) in the event of an	Yes	The robust access control and consensus mechanism will limit unauthorised transfer or access. Any unauthorised activities can also be detected and reported.

	unauthorised transfer, access or use of the non-personal data that they have shared.		
LEG-TSD-077	If TRUSTEE happens to be intermediation service according to the DGA, TRUSTEE service providers must take necessary measures to ensure an appropriate level of security for the (...) processing and transmission of non-personal data and the highest level of security for transmission of competitively sensitive information (e.g. trade secrets).	Yes	ATR enables access only to the authorised entities to record transactions. If the authorised users are compromised anyway, they cannot influence the functioning of the blockchain/ATR since a robust consensus mechanism is maintained within the network. Also, after having approval from majority of the peer nodes only a transaction can be confirmed to the ATR.
LEG-TSD-087	TRUSTEE system developers, by providing the service which involves deployment of smart contracts for others in the context of an agreement to make data available, shall ensure that the smart contract has been designed to offer a very high degree of robustness to avoid functional errors and to withstand manipulation by third parties (robustness).	Yes	ATR is a component which allows for conclusion and execution of smart contracts and, as described in Section 5, applies technical solutions of high robustness.
LEG-TSD-088	TRUSTEE system developers should ensure that a mechanism exists to terminate the continued execution of transactions: the smart contract shall include internal functions which can reset or instruct the contract to stop or interrupt the operation to avoid future (accidental) executions (safe termination and interruption). The conditions under which a smart contract could be reset or instructed to stop or interrupted, should be clearly and transparently defined. Especially, it should be assessed under which	Yes	The policies implemented within the ATR, such as endorsement policies, channel policy, Access control Policy, etc., will ensure that a transaction goes through several verification stages and also a consensus is reached. Additionally, all other requirements of Hyperledger need to be met before transactions being appended to the ledger or else the execution of the transactions will be terminated.

	conditions non-consensual termination or interruption should be permissible.		
LEG-TSD-089	TRUSTEE system developers should ensure that a smart contract is protected through and designed to offer rigorous access control mechanisms at the governance and smart contract layers (access control).	Yes	Access to the ledger will only be permitted based on the underlying access-control mechanism. E.g., extra layer of access control through Private data collection, Certificate authorities (CA), permissioned access, etc.
LEG-TSD-091	TRUSTEE system developers should ensure that a smart contract ensures the confidentiality of trade secrets.	Yes	The implementation of the off-chain storage within the smart-contract will ensure additional security.
LEG-TSD-109	TRUSTEE system developers should use only such blockchain solutions which allow to address GDPR principles, especially data minimisation, storage limitation and where necessary, the ones which will allow deletion of the data, or full anonymisation of the data that are not necessary or which erasure was requested by a data subject or a Data provider.	Yes	One of the best blockchain solutions will be utilised in terms of performance, reliability, scalability and flexibility. . It allows data deletion and solves many challenges introduced in other blockchains.
LEG-TSD-110	When personal data are stored on an immutable blockchain, the TRUSTEE system developers should ensure one of the points a-d: a) the data are either not or will no longer constitute personal data, b) that data subject (user) is considered to be the controllers of their data, c) that the household exemption defined in the GDPR applies or d) a justification for constant storage of data on the blockchain has to exist.	Yes	The personal data can be stored in the off-chain storage, Private Data Collection, of ATR which allows data deletion.

LEG-TSD-115	TRUSTEE system developers should implement solutions allowing to foresee, if a smart contract must be terminated or deactivated, a possibility to archive transactional data, the smart contract logic and code to keep the record of the operations performed on the data in the past (auditability) (data archiving and continuity).	Partially	Various measures can be adopted to back-up the recorded transactions. For example; distributed backups in multiple nodes to avoid single point of failure, setting up secure storage, ledger snapshots by copying the ledger transaction directory, etc.

12.5. KR

Req id (WP2)	Req. Description	Compliance (Y/N/Partially)	How is addressed
LEG-TSD-001	Should the TRUSTEE system developers and parties maintaining the system partners at any point handle or process any type of personal data, they would be expected to balance their legitimate interest with the fundamental rights and freedoms of the data subjects whose data they will be processing.	Yes	The KR will be implemented on private IPFS network. Thus, it offers the functionality to delete any content that contains personal data.
LEG-TSD-002	TRUSTEE system must have the fundamental rights principles at its core, by ensuring that the privacy of the individuals is respected and protected.	Yes	Being implemented on private and permissioned network the KR will ensure that the privacy is ensured.
LEG-TSD-013	With regards to the TRUSTEE user's identification data , TRUSTEE system providers will be controllers and below principles should be addressed by them:	Yes	KR will provide adequate measures to comply with the principles. For example, if sensitive information or special category data is stored, the permissioned nature of IPFS

	• Lawfulness: A legal basis in Article 6 (and where applicable an exception defined in Article 9 (2) GDPR for special categories of data) should be indicated. • Transparency: Information to the data subject (Data provider and Consumer) about processing of their data by TRUSTEE Developers and maintaining parties should be presented. • Purpose limitation: Specific, explicit and legitimate purposes and process only data for this set purpose shall be defined. • Data Minimisation: Only data which is relevant and strictly necessary to provide the service can be processed. Pseudonymisation and/or anonymisation of data should be considered wherever possible. In the case of TRUSTEE, SSI authentication should allow for protection of users' privacy. Minimisation of the amount of data collected also in the log data and not storing sensitive or critical data must be applied. • Accuracy: Technical and organisational measures shall be put in place to ensure that the data provided by users is up to date and accurate and/or allow the users to correct inaccurate data. • Storage limitation: All data which is no longer necessary for the fulfilment of the purpose shall be deleted or anonymised, unless an exception applies (e.g., scientific research). Retention periods should be associated for each category of data (including logs).	can obtain consent from the data subjects. In addition, IPFS comprises with adequate security measures and data deletion functionality. Furthermore, the unique hashing technique ensures data minimisation thus avoiding storing unnecessary data and sensitive information.

	• Confidentiality, Integrity and Data Security: Technical and organisation measures, described in the sections below and in the technical requirements shall be implemented to mitigate risks and to ensure a high level of security, confidentiality and integrity of the data.		
LEG-TSD-014	With regards to the data provided onto TRUSTEE by Data Providers, which may be considered personal, the system should allow to address these principles as well, in particular: • Lawfulness: A legal basis in Article 6 (and where applicable Article 9 GDPR) should be indicated by the data provider as the basis for providing the data onto TRUSTEE and sharing them with the data consumers, and by the Consumer for further processing for Data consumer's purposes. • Transparency: Data Provider and Consumer should be indicated as responsible for providing data subjects with the information about processing of their data on TRUSTEE and informed that the data consumers are among recipient of the data. • Purpose limitation: Data Providers should indicate limitations which are in place in the context of the purposes for which the data provided onto TRUSTEE can be processed. • Data Minimisation: Only data which is relevant to achieve the defined purpose can be processed. Pseudonymisation and/or anonymisation of data should be	Yes	The KR can be GDPR compliant, supporting the fulfilment of principles such as purpose limitation, data minimisation, accuracy and storage limitation by enabling to delete stored contents through garbage collection. Also, it can ensure the accuracy of the stored content through cryptographic hashing mechanism. The KR will be implemented on private IPFS network, hence supporting fulfilment of confidentiality and integrity principle.

	considered wherever possible. In the case of TRUSTEE, homomorphic encryption is one of the means which should contribute to data providers' and data consumers' compliance with this requirement. • Accuracy: The end users should make sure that the technical and organisational measures shall be put in place on their sides as well as on TRUSTEE allow to keep the data accurate. The users providing the data should confirm whether and how the data provided onto TRUSTEE are accurate, by providing this information in the contract management layer and in the privacy impact assessment. • Storage limitation: All data which is no longer necessary for the fulfilment of the legitimate purpose set out by the Data providers or Consumers shall be deleted or anonymised, unless an exception applies (e.g., scientific research). • Confidentiality, Integrity and Data Security: Technical and organisation measures, described in the sections below and in the technical requirements shall be implemented to mitigate risks to data processing and to ensure a high level of security, confidentiality and integrity of the data.		
LEG-TSD-012	With regard to the user data processed on TRUSTEE, TRUSTEE providers should implement solutions, such as technical and organisational measures (such as, in particular,	Yes	Being implemented on private network, the KR will ensure full privacy and will also address the GDPR requirements.

	encryption, pseudonymisation, privacy policy, and Terms of Use) which will address the requirements defined in the GDPR principles.		
LEG-TSD-021	The measures to be provided and which should be considered in the process of development and maintenance of TRUSTEE are: • Pseudonymisation and encryption of personal data – the TRUSTEE platform will employ homomorphic encryption technique; • Measures to ensure the security of IT employed in the data processing, i.e. ensuring its ongoing confidentiality, integrity, availability and resilience – to answer these requirements, TRUSTEE will employ authentication services, such as SSI and SDK; • Incident response capabilities to restore the availability and access to personal data in a timely manner – due to the federated learning implementation, this responsibility would be rather dependent on the end-users providing datasets via TRUSTEE; • A regular assessment and evaluation of the effectiveness of the implemented technical and organisational security measures. The list is non-exhaustive. The TRUSTEE providers are encouraged to provide additional security measures to ensure appropriate level of data security.	Yes	A regular assessment and evaluation can be conducted to determine the effectiveness.

LEG-TSD-023	Logging systems should be used to detect anomalies or security events.	Partially	KR does not comprise the functionalities but since it will be optimised from ATR, the ATR integration can provide such ability. Also, integration with external tools might provide such functionality.
LEG-TSD-024	Penetration testing should be considered and performed.	Partially	Benchmarking will be conducted to assess the effectiveness of KR.
LEG-TSD-025	Efficient version control system should be set up, to track different versions of the code over the time, with appropriate authentication and permission system and regular backups, as well as efficient development system.	Partially	The KR implementation does not require any coding but it can be optimised from ATR (chaincode) or the application that will be connected through API. Efficient version control will be used to backup the implementation.
LEG-TSD-026	Code quality metrics tools should be considered.	Partially	The KR implementation does not require any coding but it can be optimised from ATR (chaincode) or the application that will be connected through API.
LEG-TSD-027	Control of the use libraries and SDKs must be applied.	Partially	IPFS API provided by IPFS will be used allowing other TRUSTEE ABBs to store knowledge to KR.
LEG-TSD-028	When using free or open source software, TRUSTEE system developers should choose solutions that have an active community, good documentation and are regularly updated.	Yes	IPFS documentation will be followed.

LEG-TSD-029	When using types of solutions with commercial support, TRUSTEE system providers should contractually ensure maintenance and updates of the code during the whole lifecycle of the project.	Yes	The implementation can be updated with the latest version of IPFS during the lifecycle of the project.
LEG-TSD-030	Change of the default configurations, audit, mapping of the dependencies of the SDKs and third party libraries should be considered.	Yes	KR implementation can consider third party libraries.
LEG-TSD-031	Implementation of cryptographic algorithms and protocols by TRUSTEE system developers by their own should be avoided, and choosing cryptographic libraries which are maintained, easy to use and recognised should be encouraged.	Yes	IPFS is incorporated with well recognised cryptographic libraries that ensures enhanced security, trust and reliability.
LEG-TSD-032	Appropriate maintenance of libraries and SDKs should be considered, e.g. using dependency management system, managing updates to the dependencies, monitoring the versions of libraries and SDKs and regularly check the open source projects and their status, especially in the context of the change of domain or package ownership.	Yes	The guidelines on the maintenance details on the latest versions of libraries and other information will be followed. The maintenance can be carried out as per the documentation.
LEG-TSD-033	Quality and security of the code and its documentation, as well as the documentation of the system architecture must be ensured.	Yes	The IPFS documentation will be followed to achieve this.

LEG-TSD-034	Applications must be tested (development test, security test). Real/ production data should not be used during the development and testing phase, as using personal data for testin purposes may go beyond their initial purpose.	Partially	Although the KR implementation will not comprise development of an application, but it can be optimised through the application that is connected to KR to store content and then integrate KR to ATR. Thus, the application will go through rigorous testing phase before deployed for production.
LEG-TSD-035	When data security is comprised, TRUSTEE system providers must report the violation to the data controllers (Data Providers and Data Consumers) without undue delay if it is likely to pose a risk to the rights and freedoms of individuals.	Yes	The KR will be based on distributed network and there will be no central authority. Various approaches will be adopted to deal with security issues such as internal alerting through monitoring system, etc.
LEG-TSD-036	In case of the transfer of the data via TRUSTEE to third (non-EU) countries, appropriate safeguards must be considered and TRUSTEE should enable their implementation by the user.	Partially	The KR will be based on distributed network and it does not concern about the physical location of the nodes. The same level of security is applied to all stored contents uploaded to KR.
LEG-TSD-041	Explore whether any specific addressing of tools/considerations of allowing for the sharing of data/documents held by public sector bodies could be facilitated by TRUSTEE.	Partially	Tools can be explored (if applicable). It is to be further explored in next phases of the TRUSTEE activities.
LEG-TSD-042	TRUSTEE system developers should identify and take appropriate and proportionate technical and organizational measures to manage risks created by the security systems.	Yes	IPFS provides guidelines and best practices for securing the stored contents and nodes. It is also formed with various built-in security functionalities such as content addressing, data

			authentication, etc., which help defending security threats.
LEG-TSD-043	The level of security which should be ascertained has to be appropriate to the risks posed, and should take into account the following elements: a. the security of systems and facilities; b. incident handling; c. business continuity management; d. monitoring, auditing and testing; and d. compliance with international standards.	Yes	IPFS provides guidelines and best practices for securing the stored contents and nodes. It is also formed with various built-in security functionalities such as content addressing, data authentication, etc., which help defending security threats. Additionally, IPFS cluster, prometheus, etc., are some of the tools which can serve the requirements.
LEG-TSD-044	The security measures taken by TRUSTEE system developers shall include: (a) detection processes and procedures maintained and tested to ensure timely and adequate awareness of anomalous events; (b) processes and policies on reporting incidents and identifying weaknesses and vulnerabilities in their information systems; (c) a response in accordance with established procedures and reporting the results of the measure taken; (d) an assessment of the incident's severity, documenting knowledge from incident analysis and collection of relevant information which may serve as evidence and support a continuous improvement process.	Yes	KR will exercise all the security measures as per as its own security features and also following the guidelines set by TRUSTEE.

LEG-TSD-045	TRUSTEE system developers must fulfil the obligation of monitoring, auditing and testing, there should be policies implemented which will define supervision, inspection and verification procedures.	Yes	Hyperledger Fabric comprises with built-in security functionalities that help managing risks. Additionally, the integrated tools will provide monitoring, auditing and other functionalities. Also, Hyperledger provides guidelines/documentation and best practices for chaincode, libraries, etc. The guidelines can be followed to maintain the required obligation.
LEG-TSD-046	TRUSTEE providers may be recognised either as essential (as provider of trusted services) or as important (as research institutions) entities, therefore they should adhere, in all phases of the development of TRUSTEE and its further maintenance, to the NIS 2 Directive stemming obligations, which will be further transposed and established at the national level.	Yes	The official IPFS documentation provided by the IPFS will be followed to gain insight on the policies, guidelines, best practices, requirements, etc.
LEG-TSD-047	TRUSTEE system developers should consider and implement the requirements stemming from the NIS and NIS 2 Directive, to ensure appropriate level of security of information, especially implementation of the means such as: - policies on risk analysis and information system security, - incident handling, - business continuity , - supply chain security,	Partially	The security measures can be adopted as per the requirement and responsibility of those taking part in the implementation of the IPFS. For instance, IPFS does not provide authentication features but can be implemented at the application level that is built on top of IPFS, similarly, IPFS does not come with built-in feature for incident handling but can be leveraged by the administrators by deploying environments.

	- security in acquisition, development and maintenance of network and information system, - assessment of the effectiveness of implemented measures, - cybersecurity training and basic cyber hygiene practices, - procedures regarding the use of cryptography and, where appropriate, encryption, access control policies, asset management and HR security, - the use of multi-factor authentication or continuous authentication solutions within the entity, where appropriate.		
LEG-TSD-048	TRUSTEE system developers shall implement appropriate incident reporting procedures.	Yes	Being formed in a decentralised network, the IPFS does not have any centralised incident reporting procedure. However, various measures can be adopted by the administrators for the incident reporting procedures, such as, incident identification, evidence preservation, documentation, etc.
LEG-TSD-050	TRUSTEE providers shall thoroughly plan and implement appropriate security measures and constantly monitor, test and audit them, as well as keep the documentation describing those measures in a transparent and clear manner.	Yes	TRUSTEE provides can perform network monitoring, content integrity checks, peer monitoring, etc., to achieve this.
LEG-TSD-076	If TRUSTEE happens to be intermediation service according to the DGA,	Yes	Can achieve that through monitoring and detection,

	TRUSTEE service providers must inform data holders (data providers) in the event of an unauthorised transfer, access or use of the non-personal data that they have shared.		alerting and notifications, incident response, etc.
LEG-TSD-077	If TRUSTEE happens to be intermediation service according to the DGA, TRUSTEE service providers must take necessary measures to ensure an appropriate level of security for the (...) processing and transmission of non-personal data and the highest level of security for transmission of competitively sensitive information (e.g. trade secrets).	Yes	KR being implemented in private network enables access only to the authorised entities to store transactions. Furthermore, it comprises built-in security functionalities which ensures appropriate level of security to the stored content.

12.6. AM

Req. (WP2)	ID	Req. Description	Compliance (Yes/No/Partially)	How is addressed
LEG-TSD-001		Should the TRUSTEE system developers and parties maintaining the system at any point handle or process any type of personal data, they should balance their legitimate interest with the fundamental rights and freedoms of the data subjects whose data they will be processing.	Yes	TRUSTEE Authentication Manager will respect the fundamental rights and freedoms.
LEG-TSD-002		TRUSTEE must have the fundamental rights principles at its core, by ensuring that the privacy of	Yes	TRUSTEE Authentication Manager

	the individuals is respected and protected.		will not expose data to unauthorized users.
LEG-TSD-003	TRUSTEE must be developed as the system ensuring appropriate and secure processing of personal data. This is relevant especially for: a) the data considered as “special category data”, such as data concerning health or genetic data, as well as b) personal data concerning minors. These data will be processed by the future users of TRUSTEE (Data providers, Data consumers), especially in the HEALTH and EDUCATION domain.	Yes	TRUSTEE Authentication Manager will process personal data with the appropriate security.
LEG-TSD-005	TRUSTEE must be equipped with the technical and governance solutions which will allow for appropriate protection of the data which are or may be considered personal data, to ensure that the rights and freedoms of the data subjects (users or people whose data are accessed between users) are respected, can be exercised and will not be infringed.	Yes	TRUSTEE Authentication Manager will process personal data with the appropriate security.
LEG-TSD-006	Any website, application or server must incorporate basic state-of-the-art security rules, not only on network communications but also on authentication and infrastructure.	Yes	TRUSTEE Authentication Manager will employ SSL and TLS for secure communications.
LEG-TSD-012	With regard to the user data processed on TRUSTEE, TRUSTEE providers should implement solutions, such as technical and organisational measures (such as, in particular, encryption, pseudonymisation, privacy policy, and Terms of Use)	Yes	TRUSTEE Authentication Manager will process personal data with the appropriate security using encryption and address the

	which will address the requirements defined in the GDPR principles.		requirements defined in the GDPR.
LEG-TSD-013	With regards to the TRUSTEE user's identification data, TRUSTEE providers will be controllers and below principles should be addressed by them: • Lawfulness: A legal basis in Article 6 (and where applicable an exception defined in Article 9 (2) GDPR for special categories of data) should be indicated. • Transparency: Information to the data subject (Data provider and Data consumer) about processing of their data by TRUSTEE Developers and maintaining parties should be presented. • Purpose limitation: Specific, explicit and legitimate purposes and process only data for this set purpose shall be defined. • Data Minimisation: Only data which is relevant and strictly necessary to provide the service can be processed. Pseudonymisation and/or anonymisation of data should be considered wherever possible. In the case of TRUSTEE, SSI authentication should allow for protection of users' privacy. Minimisation of the amount of data collected also in the log data and not	Yes	TRUSTEE Authentication Manager will process personal data with the appropriate security, clearly stating the purpose of processing and giving total control of them to the user.

	storing sensitive or critical data must be applied. • Accuracy: Technical and organisational measures shall be put in place to ensure that the data provided by users is up to date and accurate and/or allow the users to correct inaccurate data. • Storage limitation: All data which is no longer necessary for the fulfilment of the purpose shall be deleted or anonymised, unless an exception applies (e.g., scientific research). Retention periods should be associated for each category of data (including logs). • Confidentiality, Integrity and Data Security: Technical and organisation measures, described in the sections below and in the technical requirements shall be implemented to mitigate risks and to ensure a high level of security, confidentiality and integrity of the data.	
LEG-TSD-014	With regards to the data provided onto TRUSTEE by data providers, which may be considered personal, the system should allow to address these principles as well, in particular: • Lawfulness: A legal basis in Article 6 (and where applicable Article 9 GDPR) should be indicated by the data provider as the basis for providing the data onto	Yes TRUSTEE Authentication Manager will process personal data with the appropriate security, clearly stating the purpose of processing and giving total control of them to the user.

TRUSTEE and sharing them with the data consumers, and by the Data consumer for further processing for Data consumer's purposes.

- Transparency: Data provider and Data consumer should be indicated as responsible for providing data subjects with the information about processing of their data on TRUSTEE and informed that the data consumers are among recipient of the data.
- Purpose limitation: Data providers should indicate limitations which are in place in the context of the purposes for which the data provided onto TRUSTEE can be processed.
- Data Minimisation: Only data which is relevant to achieve the defined purpose can be processed. Pseudonymisation and/or anonymisation of data should be considered wherever possible. In the case of TRUSTEE, homomorphic encryption is one of the means which should contribute to data providers' and data consumers' compliance with this requirement.
- Accuracy: The end users should make sure that the technical and organisational measures shall be put in place on their sides as well as on TRUSTEE allow to keep the data accurate. The users providing the data

	should confirm whether and how the data provided onto TRUSTEE are accurate, by providing this information in the contract management layer and in the privacy impact assessment. • Storage limitation: All data which is no longer necessary for the fulfilment of the legitimate purpose set out by the Data providers or Data consumers shall be deleted or anonymised, unless an exception applies (e.g., scientific research). • Confidentiality, Integrity and Data Security: Technical and organisation measures, described in the sections below and in the technical requirements shall be implemented to mitigate risks to data processing and to ensure a high level of security, confidentiality and integrity of the data.		
LEG-TSD-019	TRUSTEE must incorporate the solutions compliant with data protection requirements already in planning and development phases (privacy by design), and ensure that only personal data necessary for the purpose of providing TRUSTEE to the users are processed (privacy by default).	Yes	TRUSTEE Authentication Manager will not expose data to unauthorized users.
LEG-TSD-020	TRUSTEE must require from the end users only provision of the data necessary to fulfil the purpose of processing (i.e. authentication and necessary access information).	Yes	TRUSTEE Authentication Manager will request only data required for authentication purposed.

LEG-TSD-021	The measures to be provided and which should be considered in the process of development and maintenance of TRUSTEE are: • Pseudonymisation and encryption of personal data – the TRUSTEE platform will employ homomorphic encryption technique; • Measures to ensure the security of IT employed in the data processing, i.e. ensuing its ongoing confidentiality, integrity, availability and resilience – to answer these requirements, TRUSTEE will employ authentication services, such as SSI and SDK; • Incident response capabilities to restore the availability and access to personal data in a timely manner – due to the federated learning implementation, this responsibility would be rather dependent on the end-users providing datasets via TRUSTEE; • A regular assessment and evaluation of the effectiveness of the implemented technical and organisational security measures. The list is non-exhaustive. The TRUSTEE providers are encouraged to provide additional security measures to ensure appropriate level of data security.	Partially	TRUSTEE Authentication Manager COULD provide measures in the process of development and maintenance of TRUSTEE.

LEG-TSD-022	Appropriate, secure and differential data access management policies must be applied. Authentication should be imposed before any access to personal data.	Yes	TRUSTEE Authentication Manager will not expose data to unauthorized users.
LEG-TSD-035	When data security is comprised, TRUSTEE system providers must report the violation to the data controllers (Data providers and Data consumers) without undue delay if it is likely to pose a risk to the rights and freedoms of individuals.	Partially	TRUSTEE Authentication Manager COULD present to the data controllers any report violation occurs from the TRUSTEE system developers.
LEG-TSD-036	In case of the transfer of the data via TRUSTEE to third (non-EU) countries, appropriate safeguards must be considered and TRUSTEE should enable their implementation by the user.	Partially	TRUSTEE Authentication Manager will not use servers located in third (non-EU) countries.
LEG-TSD-041	Explore whether any specific addressing of tools/considerations of allowing for the sharing of data/documents held by public sector bodies could be facilitated by TRUSTEE.	Partially	TRUSTEE Authentication Manager COULD facilitate interoperability with external identity managers of public sector bodies.
LEG-TSD-042	TRUSTEE system developers should identify and take appropriate and proportionate technical and organizational measures to manage risks created by the security systems.	Yes	TRUSTEE Authentication Manager will be developed with the privacy and security by default principles identifying and taking appropriate security measures.
LEG-TSD-043	The level of security which should be ascertained has to be appropriate to the risks posed, and should take	Yes	TRUSTEE Authentication Manager will take into account all the security

	into account the following elements:		considerations and assess them.
	a. the security of systems and facilities; b. incident handling; c. business continuity management; d. monitoring, auditing and testing; and e. compliance with international standards.		
LEG-TSD-044	The security measures taken by TRUSTEE system developers shall include: (a) detection processes and procedures maintained and tested to ensure timely and adequate awareness of anomalous events; (b) processes and policies on reporting incidents and identifying weaknesses and vulnerabilities in their information systems; (c) a response in accordance with established procedures and reporting the results of the measure taken; (d) an assessment of the incident's severity, documenting knowledge from incident analysis and collection of relevant information which may serve as evidence and support a continuous improvement process.	Yes	TRUSTEE Authentication Manager will ensure the achievement of these security measures.
LEG-TSD-045	TRUSTEE system developers must fulfil the obligation of monitoring, auditing and testing, there should be policies implemented which will define supervision, inspection and verification procedures.	Partially	Developers of the TRUSTEE Authentication Manager COULD monitor, audit and test the AM's systems.
LEG-TSD-046	TRUSTEE providers may be either as essential (as provider of trusted services) or as important (as research institutions) entities, therefore they should adhere, in all phases of the development of	Yes	TRUSTEE Authentication Manager will be aligned with all the legal requirements.

	TRUSTEE and its further maintenance, to the NIS 2 Directive stemming obligations, which will be further transposed and established at the national level.		
LEG-TSD-047	TRUSTEE system developers should consider and implement the requirements stemming from the NIS and NIS 2 Directive, to ensure appropriate level of security of information, especially implementation of the means such as: • policies on risk analysis and information system security, • incident handling, • business continuity , • supply chain security, • security in acquisition, development and maintenance of network and information system, • assessment of the effectiveness of implemented measures, • cybersecurity training and basic cyber hygiene practices, • procedures regarding the use of cryptography and, where appropriate, encryption, access control policies, asset management and HR security, • the use of multi-factor authentication or continuous authentication solutions within the entity, where appropriate.	Yes	TRUSTEE Authentication Manager will consider and implement all the related requirements stemming from the NIS and NIS 2.
LEG-TSD-048	TRUSTEE system developers shall implement appropriate incident reporting procedures.	Yes	TRUSTEE Authentication Manager will ensure the

			achievement of these security measures.
LEG-TSD-049	TRUSTEE providers shall disclose the terms and conditions that apply to the services provided by them, as well as shall clearly describe the policies governing their provision.	Yes	TRUSTEE Authentication Manager will ensure the achievement of these security measures.
LEG-TSD-050	TRUSTEE providers shall thoroughly plan and implement appropriate security measures and constantly monitor, test and audit them, as well as keep the documentation describing those measures in a transparent and clear manner.	Partially	TRUSTEE Authentication Manager follows the security by design principle ensuring all the appropriate security measures are deployed. TRUSTEE Authentication Manager COULD monitor and audit all the security measures in place.
LEG-TSD-051	TRUSTEE system developers need to understand eIDAS framework, as they aim to build upon the eIDAS trust framework and ensure full compatibility with it in developing a self-sovereign identity management.	Yes	TRUSTEE Authentication Manager will utilise eIDAS to develop a SSI ecosystem, following the eIDAS framework.
LEG-TSD-052	In case electronic signature is applied for the log-in, the e-Signature should fulfil the requirements indicated in the Regulation.	Yes	TRUSTEE Authentication Manager will follow the appropriate regulations.
LEG-TSD-053	TRUSTEE shall fulfil the requirements imposed on the trust services and TRUSTEE system developers and Partners responsible for provision of the authentication systems should fulfil the obligations listed in Chapter III of the eIDAS.	Yes	TRUSTEE Authentication Manager will utilise eIDAS to develop a SSI ecosystem, following the eIDAS framework.

LEG-TSD-054	Identity and access management using SSI and eID (blockchain-based IAM) must support GDPR compliance, asserted and guaranteed by service provider.	Partially	Developers of the TRUSTEE Authentication Manager COULD assert and guarantee the compliance with GDPR.
LEG-TSD-055	SSI over eIDAS must enable control over geographical distribution of data (restriction to the EU or to specific countries should be possible).	Yes	TRUSTEE Authentication Manager will not use servers located to third (non-EU) countries.
LEG-TSD-056	Access information/material and credentials must be possible to be withdrawn or rotation.	Partially	TRUSTEE Authentication Manager will allow the credentials to be withdrawn (not implemented yet).
LEG-TSD-057	Logs, identity changes and access properties and rights must be GDPR compliant.	Yes	TRUSTEE Authentication Manager will follow the appropriate regulations.
LEG-TSD-058	End user (Data provider, Developer and Data consumer) data cannot be shared by default, end-users should have the possibility to opt-in and further opt-out, if they want to. Data cannot be evaluated or sold by the identity provider.	Yes	TRUSTEE Authentication Manager will not share any end user data without asking for permission or consent.
LEG-TSD-059	Identity theft must be prevented by applying control of identity export and transfer.	Partially	TRUSTEE Authentication Manager COULD try to prevent identity theft to the extent possible. (Security mechanics to prevent that will mostly be associated with the user's wallet software.)

LEG-TSD-060	Hosting model must be selected (provider-offered service / customer-managed on-premise hosting / provider- managed cloud hosting / customer-managed cloud hosting).	Yes	TRUSTEE Authentication Manager will not apply any limitation on the hosting model type.
LEG-TSD-061	Usage model of SSI over eIDAS should include rights to audit or to re-use the source code - reuse rights may be limited (e.g. as with GPL open-source license).	Yes	TRUSTEE Authentication Manager will be open source software.

13. ANNEX 3: SOCIO-ETHICAL REQUIREMENTS

The following sections is dedicated for the preliminary analysis of socio-ethical requirements presented in the D2.1 (TRUSTEE consortium, 2023) related to each component. The ID of the respective socio-ethical requirement is used as a reference to D2.1, while the requirement's description is also provided alongside whether the ABBs comply with the requirement and how. The approach presented below for addressing each socio-ethical requirement may produce additional technical, functional and non-functional requirements that will be further specified in upcoming versions of this deliverable, namely D3.2 and D3.3, as well as in D2.2. for all the ABBs included in this deliverable, namely: HEDF, AIMAAS, TAI-SDF, ATR, KR, AM.

14. ANNEX 4: TAI-SDF QUESTIONNAIRE FOR THE ASSESSMENT OF THE COMPLIANCE WITH TRUSTWORTHINESS REQUIREMENTS

TRUSTWORTHINESS OF AI

Introduction

To evaluate the Trustworthiness of AI systems in a qualitative way the checklist has been created using as reference the questions from the Assessment List of Trustworthy Artificial Intelligence (ALTAI)²⁴. The different groups of questions composing the 7 requirements of ALTAI are:

- Human agency and oversight;
- Technical robustness and safety;
- Privacy and data governance;
- Transparency;
- Diversity, non-discrimination and fairness;
- Societal and environmental wellbeing;
- Accountability.

Additionally, the initial list has been complemented with the questions based on the analysis of the AI use and data protection guidelines issued by CNIL²⁵, AEPD²⁶, UK ICO²⁷.

The resulting tables are reported below and have the following columns:

- The questions derived from ALTAI assessment list.
- The supplementary questions that stem from the guidelines analysed additionally to ALTAI, i.e. guidelines issued by CNIL²⁸, AEPD²⁹, UK ICO³⁰.
- The field to assess if the AI system proposal satisfies a specific question.
- The technical or the operational measures to put in place in the project to satisfy a specific question.
- If the measure is technical (T) or procedural (P) (this field is useful for the user to discriminate measures that need a SW implementation or measure that can be satisfied with studies or analysis);
- If the question is applicable for development (D) or/and use only (U) of the AI system (this field is used to distinguish a situation in which the user has developed the AI model from a black-box use only).

The checklist in the following tables constitute the starting point of the questionnaire about Trustworthiness: The high level questions listed here will be integrated with other lower questions in order to create a useful tool that can be used by any type of ML/AI developer, from the expert to the beginner.

²⁴ Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission, “The Assessment List for Trustworthy Artificial Intelligence (ALTAI)”, 17 July 2020, available at: <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>

²⁵ CNIL, “Self-assessment guide for artificial intelligence (AI) systems”, available at <https://www.cnil.fr/en/self-assessment-guide-artificial-intelligence-ai-systems>

²⁶ AEPD, “Audit Requirements for Personal Data Processing Activities involving AI”, available at <https://www.aepd.es/es/documento/requisitos-auditorias-tratamientos-incluyen-ia-en.pdf>; and AEPD, “GDPR compliance of processings that embed Artificial Intelligence An introduction”, available at <https://www.aepd.es/sites/default/files/2020-07/adequacion-rgpd-ia-en.pdf>

²⁷ UK ICO, “Guidance on AI and data protection”, available at <https://ico.org.uk/media/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection-2-0.pdf>, including “AI and data protection risk toolkit”, being part of the online version of the guidelines (see: <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/ai-and-data-protection-risk-toolkit/>)

²⁸ CNIL, “Self-assessment guide for artificial intelligence (AI) systems”

²⁹ AEPD, “Audit Requirements for Personal Data Processing Activities involving AI” and AEPD, “GDPR compliance of processings that embed Artificial Intelligence An introduction”

³⁰ UK ICO, “Guidance on AI and data protection” and “AI and data protection risk toolkit”

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
1) Resilience to attack and security				
1.1) Did you assess potential forms of attacks to which the AI system could be vulnerable? (ALTAI, p.9)			P	
1.1.1) Did you consider different types and natures of vulnerabilities as well as potential entry points for attacks, such as data pollution / poisoning, physical infrastructure, cyber-attacks, model evasion or model inversion? (ALTAI, p.9, UK ICO Guidance on AI and Data Protection p. 75-79)		Vulnerability analysis could be a part of the assessment proposed as deliverable in answer to question 2)	P	
1.2) Did you put measures or systems in place to ensure the integrity, robustness and resilience of the AI system against potential attacks over its lifecycle? (ALTAI, p.9, supported by AEPD Audit requirements..., p. 29 and UK ICO Guidance)		Design and develop security measures. Provide a preliminary list of defences against the following attacks: data accesses, poisoning, evasion, oracle, privacy threats.	T	
1.3) Did you put in place measures to ensure the resilience of the AI-based component against accidental or intentional manipulation? (AEPD Audit requirements..., p. 29)			T	
1.4) Did you establish procedures to monitor how the component functions, in order to be able to detect at an early stage any data leaks, unauthorized access or other security breaches? Did you implement controls over who has access to training data, training code and deployment code? (AEPD Audit requirements..., p. 29 and 30, UK ICO AI and data protection risk toolkit)			P	
1.5) Did you include a digital signature over the version of the AI-based component to guarantee its integrity? (AEPD Audit requirements..., p. 13)			T	
1.6) Did you verify how your system behaves in unexpected situations and environments? (AEPD Audit requirements..., p. 28)		Put in the validation phase specific tests to verify how your system behaves in unexpected situations and environments (part of the validation deliverable)	P	
1.7) Did you consider to what degree your system could be dual-use? If so, did you take suitable preventative measures against this case (including for instance not publishing the research or deploying the system)? (ALTAI, p.29)		To check if the system has dual use	P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
1.8) Did you red-team/pentest the model? (ALTAI, p.9)			T	
1.9) Did you consider and inform end-users of the duration of security coverage and updates? (ALTAI, p.9)			P	
1.10) Is there an automatic analysis of the logs to alert the user and/or the individual in the event of failure, abnormal operation or attack? (CNIL Self-assessment guide for AI systems: Using an AI system in production, p. 2)			T	
1.11) Could a secondary server perform processing if the main system was rendered non-operational? (CNIL Self-assessment guide for AI systems: Securing the processing, p. 2)			T	
1.12) Has a risk management system been put in place? (CNIL Self-assessment guide for AI systems: Securing the processing, p. 2; UK ICO Guidance on AI and Data Protection)			T and P	
2) Accuracy				
2.1) Did you assess what level and definition of accuracy would be required in the context of the AI system and use case? Did you review the statistical accuracy measures and how to evaluate it from the design phase? (ALTAI, p. 10, UK ICO Guidance p. 38-43, UK ICO AI and data protection risk toolkit)		Put in place an evaluation process (as part of the validation phase?) to correct unintended risks from inaccurate predictions and mistakes in classifying information in the proper categories. Address distribution drifts (covariate, label or concept drifts) and improve statistical accuracy by implementing a robust risk management framework that records classification errors over time.	P	
2.1.1) Could a low level of accuracy of the AI system result in critical, adversarial or damaging consequences? (ALTAI, p. 10)			T	
2.1.2.) Did you assess how accuracy is measured and assured? (ALTAI, p. 10, UK ICO Guidance p. 130-131)		Define a specific KPI to assess accuracy; develop a mechanism inside AI system to measure accuracy;	P	
2.1.3.) Has the accuracy of the data been verified? (CNIL Self-assessment guide for AI systems, Collecting and qualifying training data; UK ICO Guidance p. 51, 83-88; UK ICO AI and data protection risk toolkit; AEPD Audit requirements p. 21)		Accuracy assessment – make sure that the data used to train AI models are accurate, complete and representative. Inaccurate input data may cause inaccurate assumptions about data subjects that are not in line with their reasonable expectations	P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
		and may lead to negative outcomes. (UK ICO Guidance, p. 51 and 113)		
2.2) Did you put in place measures to ensure that the data used is comprehensive, of high quality, complete and representative of the environment the system will be deployed in, and up to date? (ALTAI, p. 10, UK ICO Guidance p. 66, 113)		Develop a mechanism to check dataset. When obtaining datasets from other organisations (eg as a result of a data sharing process), one need to assess the risks of bias in the same way as when collecting the data from individuals him/herself. The dataset providers and model providers should share information on data collection methodologies and any biases their datasets might have encoded. Appropriate technical approaches to mitigating possible bias should be considered, such as re-weighting, or removing the influence of protected characteristics and their proxies. (UK ICO Guidance p. 114)	T	
2.3) If an annotation method has been used, is it checked? Did you document clear criteria and lines of accountability for annotating/labelling data? (CNIL Self-assessment guide for AI systems, Collecting and qualifying training data)		Appropriately annotate datasets. The data labels should comply with data protection's accuracy principle. Clear tags of output data as inferences and predictions are required, and not claiming it to be factual. You should document criteria of accountability for labelling data to prevent data labelling that will lead to unfair outcomes for individuals. (UK ICO Guidance p. 118-119)	T and P	
2.3.1) If the annotation is performed by humans, have they been trained? (CNIL Self-assessment guide for AI systems, Collecting and qualifying training data, UK ICO Guidance p. 118-119)		Design and implement appropriate training for human reviewers.	P	
2.3.2) If yes to the above question, is the quality of their work monitored? (CNIL Self-assessment guide for AI systems, Collecting and qualifying training data)			P and T	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
2.3.4) Do you assess, document and maintain an index of data sources or features that should be not processed when making decisions about individuals to prevent an AI system producing discriminatory decisions? (UK ICO Guidance p. 119, UK ICO AI and data protection risk toolkit)		You should assess, document and record the sources or features that you do not intend to use, to prevent an AI system producing discriminatory decisions based on protected characteristics or their proxies. (UK ICO Guidance p. 119, UK ICO AI and data protection risk toolkit)	T	
2.4) Did you put in place measures to assess whether there is a need for additional data, for example to improve accuracy or to eliminate bias? (UK ICO Guidance p. 94; UK ICO AI and data protection risk toolkit)		Develop mechanisms and strategies to detect overfitting and underfitting. Examine data's context, evaluate the values you adding to your features before you feed them into the model, tune AI model hyperparameters, fit the most appropriate algorithm to the data, remove any features which may result in overfitting. (UK ICO Guidance p. 123, UK ICO AI and data protection risk toolkit)	T	
2.5) Did you put in place documented data governance practices to manage and guarantee the accuracy, integrity, veracity, topicality and adequacy of the datasets used for AI-based component's training, testing and operation? (AEPD Audit requirements p. 21, UK ICO Guidance, UK ICO AI and data protection risk toolkit)		Assess and document what data is necessary, adequate, and relevant for training and testing your AI system. Erase any data that is not needed. (UK ICO AI and data protection risk toolkit)	P	
2.5.1) Did you split the data between training, testing and validation data, where necessary? (UK ICO Guidance p. 119)			P	
2.6) Did you verify what harm would be caused if the AI system makes inaccurate predictions?		Carry out an impact assessment with socio-technical scenario in which analyses actions taken based on model's prediction	P	
2.7) Did you put in place a series of steps to monitor, document and increase the AI system's accuracy? (ALTAI, p. 10)		In the development phase, implement a strategy to achieve different levels of accuracy	P	
2.8) Did you put processes in place to ensure that the level of accuracy of the AI system to be expected by end-users and/or subjects is properly communicated, whenever required? (ALTAI, p.			P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
10, UK ICO Guidance, AEPD Audit requirements, p. 14, 16)				
2.9) Does the performance of your AI-based component take into account /implement the following: -according to the principle of data accuracy, establishment of metrics (or sets of aggregated metrics) to be used in determining the precision, accuracy and sensitivity of the component; - rate values for false positives and false negatives generated by the AI-component to improve the precision, accuracy and sensitivity of the component; -procedures to detect whether the response of the AI-based component to input data is erroneous or exceeds a certain error threshold or whether there are different error thresholds associated with different categories of individuals in the input dataset? (AEPD Audit requirements..., p. 26 and 27, UK ICO Guidance)			T	
2.10) If the processing is based on a federated learning solution, has it been checked that the data used within the centers is independent and identically distributed? CNIL Self-assessment guide for AI systems, Collecting and qualifying training data, UK ICO Guidance p. 85)			T	
2.10.1) If the data is not independent and identically distributed, what steps are taken to remedy this? (CNIL Self-assessment guide for AI systems)			P	
2.11) Are regular mechanisms in place to assess the risk of loss of quality or changes in data distribution? (CNIL Self-assessment guide for AI systems, UK ICO Guidance p. 122)			P and T	
2.12) Do you assess and document if the AI component handles data from a wide range of populations fairly and accurately? (UK ICO AI and data protection risk toolkit)		Document and assess whether your AI system can handle data from a wide range of (sub)populations fairly and accurately to detect any (sub)populations that will be unfairly treated by the AI system and to optimise the model(s). (UK ICO AI and data protection risk toolkit)	P	
3) Consistency – subcategory added based on AEPD Guidelines				
3.1) Are your AI-based component, its functioning and results yielded consistent? (AEPD Audit requirements..., p. 27, UK ICO Guidance)			T	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
3.2) Did you establish procedures and thresholds in order to control whether the yielded results diverge significantly from the expected results? (AEPD Audit requirements..., p. 27, UK ICO Guidance)			P	
3.3) Did you implement procedures to check whether the AI-based component behaves in a different manner when processing data from subjects differing in their characteristics associated to special categories of data or to proxy variables? (AEPD Audit requirements..., p. 27, UK ICO Guidance)			P	
3.4) Do you prevent discrimination and bias based on correlation between results and additional variables linked to data subjects that are normally not part of the process variable? (AEPD Audit requirements..., p. 27, UK ICO Guidance)			T	
3.5) Do you maintain quality control and matching of the data collected in the real environment with the training and validation data while using the AI system? (CNIL Self-assessment guide for AI systems: Using an AI system in production)			T	
4) Stability and robustness				
4.1) Do you monitor and ensure the stability and the robustness of your AI-based component in relation to any changes occurring in the contextual environment either internal or external to processing (e.g. new risks emerging for the rights and freedoms of data subjects, deviations in data, discrimination)? (AEPD Audit requirements..., p. 28, UK ICO Guidance)			P and T	
4.1.1) Did you document documented whether the AI component has been built with an static approach, a dynamic approach or a continuous learning approach by design? (AEPD Audit requirements..., p. 28, UK ICO Guidance)			P	
4.2) In the case of an AI system using continual learning, is a mechanisms implemented to continuously ensure the quality of the data used? (CNIL Self-assessment guide for AI systems: developing and training an algorithm)			T	
4.2.1) Where the AI-component implements a continuous learning approach, did you assess its degree of adaptability to new input data or types of data must be assessed, and put in place monitoring procedures and mechanisms in order to verify that results yielded are still valid, that AI component is still capable of obtaining new			P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
knowledge and any previous associations learned are not lost? (AEPD Audit requirements..., p. 28, UK ICO Guidance)				
4.2.2) In the continuous learning scenario: are measures taken upstream to avoid deteriorated performances, model drift or attacks aimed at influencing the results of the AI algorithm? (CNIL Self-assessment guide for AI systems: developing and training an algorithm)			T	
4.3) Are strategies such as cross-validation used? (CNIL Self-assessment guide for AI systems: developing and training an algorithm)			P and T	
4.5) Have cases of error been investigated? (CNIL Self-assessment guide for AI systems: developing and training an algorithm)			P and T	
5) Reliability and reproducibility				
5.1) Did you document why the specific AI algorithm was chosen and how implementation process looked like? (CNIL Self-assessment guide for AI systems)			P	
5.3) Has the algorithm used been tested by independent third parties? (CNIL Self-assessment guide for AI systems: developing and training an algorithm, UK ICO Guidance, p. 25, 103, 110)		Ensure that all individuals responsible for its development, testing, validation, deployment, and monitoring are adequately trained to understand the associated statistical accuracy requirements and measures;	P	
5.4) Has a comparison with similar systems been performed? (CNIL Self-assessment guide for AI systems: developing and training an algorithm)			T	
5.5) Has the hypotheses made based on the training data been discussed, clearly documented and checked against actual situation? (CNIL Self-assessment guide for AI systems)			P	
5.6) Has a study of the correlations between these particular characteristics and the rest of the training data been carried out to identify possible proxies? (CNIL Self-assessment guide for AI systems)			T	
5.7) Could the AI system cause critical, adversarial, or damaging consequences (e.g. to human safety) in case of low reliability and/or reproducibility? (ALTAI, p. 11)			P and T	
5.8) Did you put in place a well-defined strategy to monitor and test if the AI system is meeting the goals, purposes and intended applications? (ALTAI, p. 11)			P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
5.8.1) Did you test whether specific contexts or particular conditions need to be taken into account to ensure reproducibility? (ALTAI, p. 11)		To ensure reproducibility, define specific tests in the validation phase to consider specific contexts and particular conditions create “replication files” that will replicate each step of the AI system’s development process, from research and initial data collection to the results	P	
5.9) Did you put in place verification and validation as well as documentation methods to measure and ensure different aspects of the AI system’s reliability and reproducibility? (ALTAI, p. 11; AEPD Audit requirements, p. 29)		Define verification activities at design and testing phase for the reliability and reproducibility aspects (define a verification task in the validation work package)	P	
5.10) Is any monitoring scheduled for tool updates? Is there a specific protocol for changes to the AI system and updates of statistical accuracy? (CNIL Self-assessment guide for AI systems: developing and training an algorithm, UK ICO Guidance p. 43)		You should agree regular updates and reviews of statistical accuracy to guard against changing population data and concept/ model drift.	T	
5.11) Is the AI algorithm available as open source? (CNIL Self-assessment guide for AI systems: developing and training an algorithm)			T	
5.12) Is the AI algorithm in line with SOTA? (CNIL Self-assessment guide for AI systems: developing and training an algorithm)			T and P	
5.13) Did you define and justify the metrics and criteria which the verification and validation processes follow? (AEPD Audit requirements, p. 25)			T	
5.14) Did you put in place control and supervision mechanisms (e.g. testing mechanisms for the design, code, functionalities of the component, extreme test cases and boundary values, cleaning procedure to correct errors, shortcomings, inconsistencies), to ensure the compliance of the AI-based component to the goals (e.g. preserving information integrity) and purposes set? (AEPD Audit requirements, p. 25)			P and T	
5.15) Are the code changes versioned? (CNIL Self-assessment guide for AI systems: Securing the processing)			T	
5.16) Has a continuous integration and distribution process, usually referred to as the CI/CD pipeline, been implemented? (CNIL Self-			T	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
assessment guide for AI systems: Securing the processing)				
5.17) Did you put in place processes to describe when an AI system fails in certain types of settings? (ALTAI, p. 11)		To carry out specific tests with different type of settings and report the results also in case of AI system failure	P	
5.18) Did you clearly document and operationalise these processes for the testing and verification of the reliability of AI systems? (ALTAI, p. 11)		As result of verification task for the reliability	P	
5.19) Did you define a testing strategy and a complete testing plan to assess the correction of the AI component both from structural and functional terms, for early identification and remedy of defects in requirements or design, incorrect specifications or deviations from applicable criteria during development? (AEPD Audit requirements, p. 24-26, UK ICO AI and data protection risk toolkit)			T and P	
5.20) Did you put in place a proper procedure for handling the cases where the AI system yields results with a low confidence score? (ALTAI, p. 11)			P	
5.21) Did you consider potential negative consequences from the AI system learning novel or unusual methods to score well on its objective function? (ALTAI, p. 11)			T and P	
5.22) Did you establish mechanisms of communication to assure (end-)users of the system's reliability? (ALTAI, p. 11)		Mechanisms to log AI actions and to audit logged events (e.g. anomalies, failure, model requests, etc.) will be developed?	P	
5.23) Did you put in place mechanisms to limit the context of use of the AI-based component within your organisation? (AEPD Audit requirements, p. 15)			P	
5.24) Did you clearly define the requirements applicable to human operators in charge of supervising and interpreting the operation of the AI-based component? (AEPD Audit requirements, p. 15-16, UK ICO Guidance, p. 98, 103, CNIL Self-assessment guide for AI systems: Using an AI system in production)			P	
6) Unfair bias avoidance				
6.1) Has a bias study been carried out? Did you document and assess different ways the AI system/model could result in unfairness, and did you include appropriate technical and organisational measures you will use to mitigate or manage those risks on a continual basis?		Documentation of an assessment of the different ways your AI system could result in unfairness, including information about technical and	P and T	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
(CNIL Self-assessment guide for AI systems: Collecting and qualifying training data; UK ICO AI and data protection risk toolkit)		organizational measures used to mitigate those risks. Consider processes to report security breaches, and who is responsible for handling and managing them as part of an AI incident response plan.		
6.2) Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design, and defined the approach to discrimination mitigation? (ALTAI, p. 16, AEPD Audit requirements p. 24; see also CNIL Self-assessment guide for AI systems: Collecting and qualifying training data, UK ICO AI and data protection risk toolkit, UK ICO Guidance, p. 105-134)		A conceptual assessment of bias will be carried out as part of the verification task. On the basis of assessment of bias a strategy will be defined to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design	P	
6.3) Is there a possibility of bias due to the data collection method used or specific data collection conditions? Did you assess the risks of such bias in the data collected by you or obtained from a third party? (CNIL Self-assessment guide for AI systems: Collecting and qualifying training data, UK ICO Guidance p.114)		When obtaining datasets from other organisations (eg as a result of a data sharing process), one need to assess the risks of bias in the same way as when collecting the data from individuals him/herself. The information on data collection methodologies and any biases the datasets might have encoded should be made available. (UK ICO Guidance p. 114)	P	
6.4) Did you assess and acknowledge the possible limitations stemming from the composition of the used data sets? (ALTAI, p. 16)		the conceptual assessment of bias (as part of the verification task) should include the analysis of limitations stemming from the composition of the used data sets	P	
6.5) If a publicly accessible database has been used, has it been studied, in particular for the presence of any bias? (CNIL Self-assessment guide for AI systems: Collecting and qualifying training data; see also UK ICO Guidance, p. 114)			T	
6.6) Did you consider diversity and representativeness of users in the data? Did you test for specific populations or problematic use cases? (ALTAI, p. 16)		The selection of the dataset will consider the diversity and the representativeness of users in the data. The validation phase will test AI for specific populations or problematic use cases	P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
6.7) Did you conduct necessary analyses, at the beginning and in each iteration of the learning process, as well as on the samples used to train the AI model, to control that the final dataset is representative of the context to which the AI-based component is oriented? (AEPD Audit requirements... p. 22; see also UK ICO Guidance, p. 66, 113, 114, 116, 119)		You should also detect any biases that may be encoded into the data you collect or procure. Assess and document which protected characteristics data you will collect for bias analysis to detect and correct an AI system exhibiting bias. Make sure that the algorithm incorporates regularisation terms, imposes algorithmic fairness constraints or incorporates selected metrics into its objective function. (UK ICO Guidance, p. 119, UK ICO AI and data protection toolkit)	T	
6.8) Did you research and use available technical tools, being state-of-the-art, to improve your understanding of the data, model and performance? (ALTAI, p. 16)		<i>Option 1</i>) the AI system will use existing explainable AI libraries or framework or <i>Option 2</i>) the project will design and develop a specific module to be integrated in the AI system to improve the user understanding of the data, model and performance	T	
6.9) Did you put in place processes to test and monitor for potential biases during the development, deployment and use phase of the system, i.e. the whole lifecycle of the AI system? (ALTAI, p. 16)		Monitoring for potential biases during the development, deployment and use phase of the system should be put in place. Assess and document which protected characteristics data will be collected for bias analysis to detect and correct an AI system exhibiting bias. (UK ICO AI and data protection risk toolkit)	P	
6.10) Did you put in place measures, where appropriate, to incorporate additional data to improve the precision of the model and remove bias? (AEPD Audit requirements... p. 24, 26, see also UK ICO Guidance)			T	
6.11) Did you implement measures to verify that there are no historical biases in the input data used for training the AI-based component /			T	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
system and where this is the case, a different and bias-free dataset is chosen for training or the training data is made suitable for use through cleaning and scrubbing? (AEPD Audit requirements... p. 24; UK ICO Guidance p.113, 118, 122, 123)				
6.12) Depending on the use case, did you ensure a mechanism that allows to flag issues related to bias, discrimination or poor performance of the AI system? (ALTAI, p. 16, see also UK ICO Guidance p. 103)		The developed AI system will include mechanisms that allow to flag issues related to bias, discrimination or poor performance of the AI system	T	
6.12.1) Did you establish clear steps and ways of communicating on how and to whom such issues can be raised? (ALTAI, p. 16)			P	
6.12.2) Did you consider others, potentially directly or indirectly affected by the AI system, in addition to the (end)-users? (ALTAI, p. 16)		Carry out an impact assessment with socio-technical scenario that considers all actors that can be potentially directly and indirectly affected by the AI system (not only the end users)	P	
6.13) Did you implement any human supervision/revision mechanisms to control and ensure that results of the component / model are bias-free? (AEPD Audit requirements... p. 24; see also UK Guidance p. 101, 130, 131)		Design and implement appropriate training for human reviewers. Give staff the appropriate authority, incentives and support to address or escalate individuals' concerns and, if necessary, override the AI system's decision. (UK ICO AI and data protection risk toolkit)	P	
6.14) Did you implement mechanisms to enable individuals to request human intervention, provide feedback or challenge the results obtained by means of automated decision-making algorithms? (AEPD Audit requirements... p. 24; see also UK Guidance and UK ICO AI and data protection risk toolkit)		A transparent and simple-to-use mechanism should be implemented, allowing impacted individuals and groups to contest decisions they deem unfair and to comment on the explanations they receive. Recording and monitoring the number of individuals' challenges to the AI system's decision on fairness grounds is recommended. This helps in evaluating the data protection compliance risks. (UK ICO Guidance)	T	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
6.15) Did you assess whether there is any possible decision variability that can occur under the same conditions? (ALTAI, p. 32)		Option 1) Carry out a validation campaign to assess whether there is any possible decision variability that can occur under the same conditions Option 2) carry out an impact assessment with socio-technical scenario to assess whether there is any possible decision variability that can occur under the same conditions	P	
6.15.1) If so, did you consider what the possible causes of this could be? (ALTAI, p.32)		Option 1) the test will consider the possible causes of variability Option 2) the impact assessment will consider the possible causes of variability	P	
6.15.2) In case of variability, did you establish a measurement or assessment mechanism of the potential impact of such variability on fundamental rights? (ALTAI, p.32)		In case of variability, the project will establish a measurement of the potential impact of such variability on fundamental rights	P	
6.15.3) Do you test whether your AI system produces similar outcomes for entities who have different protected characteristics? (UK ICO AI and data protection risk toolkit)		Measure different types of error (eg false positives, false negatives, etc). Consider whether the testing dataset is adequate. Record any limitations of the model in the context of statistical inaccuracies. Document the limitations in a model card. (UK ICO AI and data protection risk toolkit)	T	
6.16) Did you ensure an adequate working definition of “fairness” that you apply in designing AI systems? (ALTAI, p. 17, see also UK ICO Guidance, p. 55-56, 111)		The project will consider the definition of fairness as the “non existence of bias” and will be assessed with the following metrics: - equalized odds; - equalized opportunity; - demographic parity; - fairness through awareness/unawareness. (see: UK ICO Guidance)	P	
6.16.1) Is your definition commonly used? Did you consider other definitions before choosing this one? (ALTAI, p. 17)		The definition has been given by the German BSI (federal office for information security) in the document “AI Cloud	P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
		Service Compliance Criteria Catalogue (AIC4)". Also the definition of the principle of fairness given by the European Commission High Level Expert Group on Artificial Intelligence		
6.16.2) Did you ensure a quantitative analysis or metrics to measure and test the applied definition of fairness? (ALTAI, p. 17)			T	
6.16.3) Did you establish mechanisms to ensure fairness in your AI systems? Did you consider other potential mechanisms? (ALTAI, p. 17; UK ICO Guidance – section on fairness)		On the basis of the project, specific fairness mechanisms will be implemented that can be classified in three categories: -Preprocessing that aims at building a new representation of input data by removing the information correlated to the sensitive attribute -Optimization at training time provides good performance on accuracy and fairness measure and ensures higher flexibility in optimizing the trade-off between these factors -Post processing are focused on editing the posteriors to satisfy the fairness constraints	T	
7) Minimising and reporting negative Impact				
7.1) Does the AI system replace another type of system for the task it is assigned? (CNIL Self-assessment guide for AI systems: Asking the right questions before using an artificial intelligence System)			T and P	
7.1.1) Does the AI system have a significant advantage compared to other available solutions? (CNIL Self-assessment guide for AI systems: Asking the right questions before using an artificial intelligence System)			P	
7.1.2) Does this significant advantage outweigh the potential additional risks associated with the use of an AI system? (CNIL Self-assessment guide for AI systems: Asking the right questions before using an artificial intelligence System)			P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
7.1.3) Considering the answers to the two above questions, does the use of an AI system to achieve the identified purpose seem proportionate and necessary? (CNIL Self-assessment guide for AI systems: Asking the right questions before using an artificial intelligence System)			P	
7.2) Did you carry out a risk or impact assessment of the AI system, which takes into account different stakeholders that are (in)directly affected? (ALTAI, p.33)		Carry out an impact assessment of the AI system, which takes into account different stakeholders that are (in)directly affected	P	
7.2.1) Did you assess whether your AI system is a high risk AI system? (AI Act, Article 6)			P	
7.3) Did you provide training and education to help developing accountability practices such as risk assessment? (ALTAI, p. 8, UK ICO Guidance p. 13, 14, 19, 96, 101, 103, 116, 119)		Training task on AI system will also help developing accountability practices. People assigned to provide human oversight should remain engaged, critical and able to challenge the system's outputs wherever appropriate and once the system is deployed.	P	
7.3.1) Which employees or branches of the team are involved? Does it go beyond the development phase? (ALTAI, p. 22)		Training task on AI accountability practices will also involve end users of the AI system	P	
7.3.2) Do these trainings also teach the potential legal framework applicable to the AI system? (ALTAI, p. 22)		Training task on AI accountability practices will also include legal framework applicable to the AI system	P	
7.3.4) Did you consider establishing an 'ethical AI review board' or a similar mechanism to discuss overall accountability and ethics practices, including potentially unclear grey areas? (ALTAI, p. 22)		To consider in the management structure of the AI model application an ethical AI review board	P	
7.4) Did you foresee any kind of external guidance or put in place auditing processes to oversee ethics and accountability, in addition to internal initiatives? (ALTAI, p. 22)		To include auditing processes to oversee ethics and accountability. Clear audit trails are also relevant to satisfy accountability and documentation requirements.	P	
7.5) Did you establish processes for third parties (e.g. suppliers, consumers, distributors/vendors) or workers to report potential vulnerabilities, risks or biases in the AI system? Does it go beyond development phase? (ALTAI, p. 22)		Establish a process for third parties (e.g. suppliers, consumers, distributors/vendors) or workers to report potential vulnerabilities, risks or biases in the AI system	P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
		(this could be part of the validation report) (ALTAI, p. 22)		
Socially robust, in that they duly consider the context and environment in which they operate				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall project or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
8) Fundamental rights				
8.1) Did you carry out a fundamental rights impact assessment where there could be a negative impact on fundamental rights? Did you identify and document potential trade-offs made between the different principles and rights? (ALTAI p. 5, 6, 21, 22)		Carry out a fundamental rights impact assessment where there could be a negative impact on fundamental rights.	P	
8.2) Does the AI system guide, interact or take decisions by human (end) users (e.g. recommended actions or decisions to take, presenting of options)? (ALTAI, p. 7)		Identify during the proposal preparation if the AI system interact with decisions by human (end) users	P	
8.2.1) Could the AI system affect human autonomy by interfering with the (end) user's decision-making process in an unintended or undesirable way? (ALTAI, p. 7, UK ICO Guidance p. 100-102, p. 130)		Identify during the proposal preparation if the AI system affect human autonomy by interfering with the (end) user's decision-making process in an unintended way	P	
8.2.2) Could the AI system affect human autonomy by generating over-reliance by end-users? Do you have procedure to avoid it? (ALTAI, p. 7)			P	
8.2.3) Could the AI system generate confusion for some or all end-users or subjects on whether a decision, content, advice or outcome is the result of an algorithmic decision? (ALTAI, p. 7)			T and P	
8.3) Did you consider whether the AI system should communicate to (end) users that a decision, content, advice or outcome is the result of an algorithmic decision? Are end-users or other subjects adequately made aware that a decision, content, advice or outcome is the result of an algorithmic decision? (ALTAI, p. 7)		Implement a mechanism inside the AI system that communicate to (end) users that a decision, content, advice or outcome is the result of an algorithmic decision	T	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
8.4) In case of a chatbot or other conversational system, are the human end users made aware that they are interacting with an AI system / non-human agent? (ALTAI, p. 15)		In case of a chatbot or other conversational system, the human end users will be aware that they are interacting with a non-human agent	T	
8.5) Did you analyse and if so, how do you manage the risks to the rights and freedoms of data subjects introduced by using an AI-based component in data processing? (AEPD Audit requirements p. 16, 26 and 29; ALTAI p. 5 and 6; UK ICO Guidance p. 18, 37, 47, 54, 92; CNIL Self-assessment guide for AI systems: Ensuring individuals can fully exercise their rights)			P and T	
8.6) Did you define the requirements concerning data protection and security at the origin, while taking into account the available standards and best practices? (AEPD Audit requirements p. 29; UK ICO Guidance p. 50, 72-81; CNIL Self-assessment guide for AI systems: Securing the processing)			P	
8.6.1) Do the results of your risk analysis feed into the determination of the security and privacy requirements for the AI-based component? (AEPD Audit requirements p. 29; UK ICO Guidance p. 18, 37, 47, 54, 92; CNIL Self-assessment guide for AI systems: Securing the processing)			P	
9) Human agency				
9.1) Is the AI system implemented in work and labour process? If so, did you consider the task allocation between the AI system and humans for meaningful interactions and appropriate human oversight and control? (ALTAI p.28); see also UK ICO Guidance, p. 98, 101, 108-109, CNIL Self-assessment guide for AI systems: Using an AI system in production)		In case of the AI system is implemented in work and labour process, the project should consider the task allocation between the AI system and humans for meaningful interactions and appropriate human oversight and control	P	
9.1.1) Does the AI system enhance or augment human capabilities? (ALTAI, p. 20)		Identify during the proposal preparation if the AI system enhance or augment human capabilities	P	
9.2.2) Did you take safeguards to prevent overconfidence in or overreliance on the AI system for work processes? (ALTAI, p. 7)		In the design of AI system, implement specific mechanism to prevent user overconfidence in or overreliance on the AI system	T and P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
9.3) Does the AI system risk creating human attachment, stimulating addictive behaviour, or manipulating user behaviour? (ALTAI, p. 8)			P	
9.4) Did you take measures to deal with possible negative consequences for end-users or subjects in case they develop a disproportionate attachment to the AI System? (ALTAI, p. 8)			P	
9.5) Did you take measures to minimise the risk of addiction? (ALTAI, p. 8)			T and P	
9.6) Did you take measures to mitigate the risk of manipulation? (ALTAI, p. 8)			T and P	
10) Human oversight				
10.1) Did you consider the appropriate level of human control for the particular AI system and use case? (ALTAI, p. 8)		In the design of AI system, consider the appropriate level of human control for the particular AI system and use case	T	
10.1.1) Can you describe the level of human control or involvement? Is the AI system a self-learning or autonomous system? Is it overseen by a “human-in-the-loop”? Is it overseen by a “human-on-the-loop”? Is it overseen by a “Human-in-command”? (ALTAI, p. 8)		Describe the level of human control or involvement with respect to AI system	P	
10.1.2) Who is the “human in control” and what are the moments or tools for human intervention? (ALTAI, p. 8; CNIL Self-assessment guide for AI systems: Using an AI system in production)		Describe who is the “human in control” and what are the moments or tools for human intervention?	P	
10.1.3) Did you put in place mechanisms and measures to ensure human control or oversight? Did you train the humans on how to exercise oversight? (ALTAI, p. 8; UK ICO Guidance, p. 66, 98, 99, 103; CNIL Self-assessment guide for AI systems: Developing and training an algorithm)		In the design phase of AI system implement mechanism to ensure human control or oversight	T	
10.1.4) Did you take any measures to enable audit and to remedy issues related to governing AI autonomy? (ALTAI, p. 21; AEPD Audit requirements p. 15-16, 18, 19, 22)		In the design phase of AI system implement mechanism to enable audit and to remedy issues related to governing AI autonomy	T	
10.2) Is there a self-learning or autonomous AI system or use case? If so, did you put in place more specific mechanisms of control and oversight? (ALTAI, p. 8)		If there is a self learning or autonomous AI system, implement mechanisms of control and oversight	T	
10.2.1) Which detection and response mechanisms did you establish to assess whether something could go wrong and in case of any undesirable adverse effects of the AI system for end-users or subjects? (ALTAI, p. 8)		In the design phase of AI system, implement detection and response mechanism to assess	T	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
		whether something could go wrong		
10.2.2) Did you ensure a stop button or procedure to safely abort an operation where needed? Does this procedure abort the process entirely, in part, or delegate control to a human? (ALTAI, p. 8)		In case of autonomous AI system design a stop button or procedure to safely abort an operation where needed	T	
10.3) Have the necessary human and material resources been planned? Are they clearly identified? (CNIL Self-assessment guide for AI systems: Using an AI system in production)			P	
10.4) Has the identification of cases where human intervention is necessary been put in place? (CNIL Self-assessment guide for AI systems: Using an AI system in production)			T and P	
10.5) Is an action logging system for human intervention / human review in place? (CNIL Self-assessment guide for AI systems: Using an AI system in production)			T	
11) Respect for Privacy and Data Protection when using AI				
11.1) Depending on the use case, did you establish a mechanism allowing others to flag issues related to privacy or data protection in the AI system's processes of data collection (for training and operation) and data processing? (ALTAI, p. 12)		Design a mechanism allowing users to flag issues related to privacy or data protection in the AI system's processes of data collection and data processing?	T	
11.2.) Did you establish mechanisms to supervise data collection, processing, storage and use processes? (ALTAI, p. 12 and 13, AEPD Audit requirements, p. 21-22, CNIL Self-assessment guide for AI systems; UK ICO Guidance)			T	
11.3) Did you assess the type and scope of data in your data sets (for example whether they contain personal data)? (ALTAI, p. 12; CNIL Self-assessment guide for AI systems: Asking the right questions before using an artificial intelligence system & Collecting and qualifying training data; AEPD Audit requirements, p. 21-24; UK ICO Guidance)		As part of the impact assessment, assess the type and scope of data in your data sets	P	
11.4) Did you document the structure and properties of your datasets (e.g. extension of used data, any classifying variables; where the dataset contains personal data, the number of data subjects, special categories of personal data (in line with the GDPR))? (AEPD Audit requirements, p. 23)			T and P	
11.5) Do you envision automated-decision making or profiling with the use of AI system? (GDPR Art. 22, CNIL Self-assessment guide for AI systems: Asking the right questions before using an artificial intelligence system &		If this is envisioned, additional requirements from Article 22 GDPR need to be fulfilled.	P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
Ensuring individuals can fully exercise their rights; AEPD Audit requirements p. 12, UK ICO Guidance p. 51-52, 64, 69-71)		You must tell people whose data you are processing that you are doing so for automated decision-making and give them “meaningful information about the logic involved, as well as the significance and the envisaged consequences” of the processing for them. Under Article 15 (2)(h) you must also tell them about this if they submit a subject access request. Additionally, you should apply other suitable safeguards when processing personal data to make automated decision with legal or similar effects. The safeguards include the right for individuals to: a. obtain human intervention; b. express their point of view; c. contest the decision made about them. (UK ICO Guidance, p. 94)		
11.6) Are individuals who interact with the AI system or about whom an automated decision will be made identified and aware of the conditions? (GDPR Art. 22, CNIL Self-assessment guide for AI systems: Collecting and qualifying training data & Using an AI system in production; AEPD Audit requirements p. 12, UK ICO Guidance p. 94-95)			P	
11.5) Do you collect the data specifically for training of a specific AI system? (CNIL Self-assessment guide for AI systems: Collecting and qualifying training data)			P	
11.6) If you reuse already existing data for the purposes of training and development of the AI system, has the database been compiled and data collected in accordance with data protection regulations? (CNIL Self-assessment guide for AI systems: Collecting and qualifying training data)			P	
11.7) Do you monitor the compliance of training data processing with data protection			P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
requirements? (ALTAI, p. 13; CNIL Self-assessment guide for AI systems: Collecting and qualifying training data; UK ICO Guidance)				
11.8) Did you consider ways to develop the AI system or train the model without or with minimal use of potentially personal data (including special categories/sensitive data)? (ALTAI, p. 13; CNIL Self-assessment guide for AI systems: Collecting and qualifying training data; see also AEPD Audit requirements p. 16; UK ICO Guidance)		In the conceptual phase of the AI system, consider ways to develop the AI system or train the model without or with minimal use of potentially sensitive or personal data	T	
11.9) Does the way in which the datasets used for training are produced and used meet the data minimisation principle? Did you cleanse or sanitize your data collected and used for the training of the AI-based component from all information which is not strictly necessary to train the model, in compliance with the data minimization principle? (CNIL Self-assessment guide for AI systems: Collecting and qualifying training data; AEPD Audit requirements p. 23; UK ICO Guidance p. 83-84)		Assessment if the data is accurate, adequate, relevant and limited to the purposes of processing. Apply data minimisation principle and privacy measures in the different stages of developing the AI-based component (incl. strategies, such as data hiding, separation, anonymization, and pseudonymisation)	T and P	
11.11) Could the collection of some values be avoided if they prove not to be useful or necessary for learning, especially if they include sensitive/special categories of data? If the collection of these values cannot be avoided, will the data be removed or hidden? (CNIL Self-assessment guide for AI systems: Collecting and qualifying training data; see also UK ICO Guidelines p. 55-56)			T and P	
11.12) Did you document your data sources and implement an information mechanism concerning the identified characteristics of the data used to train the AI-based component and the justifications for such use? (AEPD Audit requirements.. p. 22 and 27; UK ICO Guidance)			P	
11.13) Did you justify the collection and use of personal data when such data are not necessary in the training stage, in order to test the model behaviour in the subsequent stages of component verification and validation? (AEPD Audit requirements p. 22; CNIL Self-assessment guide for AI systems: Collecting and qualifying training data; see also UK ICO Guidelines)			P	
11.15) Did you document and justify the process of selecting the data sources used for training the AI-based component? (justification including the			P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
legal grounds behind using personal data in the various life-cycle stages of the AI-based component as well as the justification for the use of processing any sensitive personal data) (AEPD Audit requirements p. 22)				
11.16) Does the processing directly or indirectly target vulnerable individuals (children, patients, employees, etc)? (CNIL Self-assessment guide for AI systems: Asking the right questions before using an artificial intelligence system)			T and P	
11.17) Did you identify the origin and the context of the data sources used for training and validating the AI-model? (AEPD Audit requirements p. 22; CNIL Self-assessment guide for AI systems: Collecting and qualifying training data; UK ICO Guidance p. 74, p. 114)			P	
11.18) Do you, and if so, how do you make available information regarding metadata of the AI-based component, the logic behind it and the potential consequences of the use of the model available and accessible to data subjects? (ALTAI p. 15; AI Act, Art. 13; AEPD Audit requirements p. 14; UK ICO Guidance p. 94;)			P	
11.19) Did you inform your data subjects of the logic involved behind the AI-based component, the manner in which processing is performed, and of the envisaged consequences (ALTAI p. 15; AI Act, Art. 13; AEPD Audit requirements p. 18, UK ICO Guidance p. 94, ?			P	
11.20) Did you assess the use of the AI-based component / system against any possible options for any processing of personal data, in consideration of the fundamental rights of data subjects? (ALTAI, p. 5-6; AEPD Audit requirements p. 14, 30; UK ICO Guidance; UK ICO AI and data protection toolkit)			P	
11.21) Did you put in place procedure to evaluate the efficiency and adequateness of the functioning and results of the AI-based component against any more thoroughly tested components, which use stricter data minimisation criteria (especially such components making less use of special categories of personal data) or presenting less threats/risks for the fundamental rights of data subjects? (AEPD Audit requirements, p. 16)			P	
11.22) Do you store the data used for training the AI system / AI-based component as well as the data generated by this system / component pursuant to the storage limitation principle, where the data is stored no longer than needed for the intended purposes? (AEPD Audit		Regular internal assessment which data are relevant to be stored and why it is required. Assess which privacy –enhancing	T and P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
requirements, p. 17; ALTAI p. 12; UK ICO Guidance p. 50, 113)		technologies can be applied here.		
11.23) Did you define the legal grounds to process the data in the context of development and deployment of the AI model? (ALTAI p. 5-6, 12; AEPD Audit requirements, p. 15, 17, 22; UK ICO Guidance 29-37; CNIL Self-assessment guide for AI systems: Collecting and qualifying training data; UK ICO AI and data protection toolkit)			P	
11.23.1) Did you define the legal grounds to store personal data used by the AI-based component for a period of time that exceeds the period established for processing purposes? (AEPD Audit requirements, p. 17; UK ICO Guidance p. 50; UK ICO AI and data protection toolkit)			P	
11.24) In line with the above-question, did you define and if so, do you implement a storage policy and further privacy strategies (e.g. minimisation, hiding, separation or abstraction) for the personal data? (AEPD Audit requirements, p. 17; UK ICO Guidance p. 50; UK ICO AI and data protection toolkit)			P	
11.25) Did you define and put in place the necessary and appropriate technical and organizational measures for the storage of personal data? (AEPD Audit requirements, p. 18; UK ICO Guidance p. 51, 72-88; UK ICO AI and data protection toolkit)			P	
11.26) Did you define the time limits for erasure of stored personal data? (AEPD Audit requirements, p. 18; UK ICO Guidance p. 50, UK ICO AI and data protection toolkit)			P	
11.27) Did you put in place procedures to verify the implementation of the storage periods? (AEPD Audit requirements, p. 18)			P	
11.28) Did you determine the life cycle stages of the AI-based component in which the processed personal data are to be stored? (AEPD Audit requirements, p. 17; UK ICO Guidance p. 50, UK ICO AI and data protection toolkit)			P	
11.29) Is the objective (purpose) of the processing based on the use of an AI system clearly defined? (CNIL Self-assessment guide for AI systems: Asking the right questions before using an artificial intelligence system; UK ICO Guidance p. 48, 49, 63, 64; UK ICO AI and data protection toolkit)			P	
11.30) Will the processing result in legal, financial or physical consequences for the health, social status or safety of the individuals directly or indirectly targeted by the AI systems or			T and P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
similarly affect them? (CNIL Self-assessment guide for AI systems: Asking the right questions before using an artificial intelligence system)				
11.31) Did you build in mechanisms for notice and control over personal data depending on the use case (such as valid consent and possibility to revoke, when applicable, right to object or right to be forgotten)? (UK ICO Guidance, p. 89-104, AEPD Audit requirements p. 14, 17; CNIL Self-assessment guide for AI systems: Ensuring individuals can fully exercise their rights)		Design mechanisms for notice and control over personal data depending on the use case (such as valid consent and possibility to revoke, when applicable). Document how you will facilitate individual rights requests throughout the lifecycle of your AI system where personal data will be processed.	T	
11.32) Do you fulfill your information obligation to data subjects, whose data are to be processed by the AI-based component? This is regardless of how data was collected from data subjects (i.e. directly from them or obtained otherwise), unless it is not possible to identify an individual directly or indirectly. (GDPR, Art. 14; UK ICO Guidance, p. 92, UK ICO AI and data protection toolkit; AEPD Audit requirements p. 17; CNIL Self-assessment guide for AI systems: Ensuring individuals can fully exercise their rights)			P	
11.33) Have you given the users and operators of the AI-based component sufficient information and made them aware of their duties and responsibilities regarding security and protection of data and data subjects' rights and freedoms? (AEPD Audit requirements p. 30; CNIL Self-assessment guide for AI systems: Developing and training an algorithm; UK ICO Guidance p. 96; UK ICO AI and data protection toolkit)			P	
11.34) Did you identify the categories of data subjects that are and/or might be affected from the development of an AI-based component and as a result of its use (both short and long term consequences) in the data processing procedure? (AEPD Audit requirements p. 18; UK ICO Guidance p. 26; UK ICO AI and data protection toolkit)			P	
11.35) Did you identify and record any potential recipients or categories of recipients, including those in third countries, of personal data processed by the AI-based component? Are the transfers of personal data by controllers to any recipients in third countries justified and monitored? (AEPD Audit requirements p. 17; UK ICO Guidance p. 79)		Document clear audit trails of how personal data is moved and stored from one location to another during the training and testing phase. (UK ICO AI and data protection toolkit)	P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
11.36) Are sector-specific regulations also considered and involved in the policies regulating the development of the AI-based component? (AEPD Audit requirements p. 18-19; ALTAI p. 13; CNIL Self-assessment guide for AI systems: Achieving compliance; UK ICO Guidance p. 45, 104)			P	
11.37) Did you take measures to enhance privacy-by-design, such as via encryption, pseudonymisation, anonymisation and aggregation? (ALTAI, p. 12; GDPR Art. 25 and 32; AEPD Audit requirements p. 23, 29-30; UK ICO Guidance p. 84-88; CNIL Self-assessment guide for AI systems: Collecting and qualifying training data)		Design security mechanism to enhance privacy such as encryption and anonymization	T	
11.38) Did you assess the degree of anonymisation and possible risk of re-identification? (AEPD Audit requirements p. 24; CNIL Self-assessment guide for AI systems: Collecting and qualifying training data; UK ICO Guidance p. 84-88)			T and P	
11.39) Where a Data Privacy Officer (DPO) exists, did you involve this person at an early stage in the process? (ALTAI, p. 12; AEPD Audit requirements p. 19; UK ICO Guidance p. 19)		Involve the DPO at an early stage in the process (proposal preparation or from the beginning of the project execution)	P	
11.40) Is the DPO assisted and advised by experts on specific matters relating to the AI component being audited? (AEPD Audit requirements p. 19)			P	
11.41) Did the DPO to take an active role in providing advice and participating in the selection, design, and development of the AI component used in personal data processing? (AEPD Audit requirements p. 19)			P	
11.42) Did you identify the parties exercising control over the life-cycle of the AI-based component? (inc. defining the organisational and/or contractual structures between the parties and the distribution of tasks and responsibilities for e.g. maintenance, updating and minimising system privacy issues) (AEPD Audit requirements p. 13; ALTAI p. 21-22)		It must be documented. Please also consider supporting with relevant documentation outlining the roles of the controllers and processors of the relevant processing activities to specify the distribution of roles and responsibilities for the protection of personal data, where personal data is involved.	P	
11.43) Did you put in place mechanisms to keep a record of the processing activities conducted by the controllers and processors? (AEPD Audit requirements p. 14; UK ICO Guidance p. 74)			T and P	
11.44) Did you document the legal, social, economic, organizational, technical, scientific or			P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
other contexts related to the inclusion of the AI-based component in the processing? (AEPD Audit requirements p. 15)				
11.45) Before processing personal data for the development of the AI system, did you conduct a data protection impact assessment? Has the risk analysis taken into account the specific attack patterns of AI algorithms? (ALTAI, p. 12, AEPD Audit requirements p. 16; CNIL Self-assessment guide for AI systems: Achieving compliance; UK ICO Guidance & UK ICO AI and data protection toolkit)			P	
11.46) Did you consider the privacy and data protection implications of data collected, generated or processed over the course of the AI system's life cycle? (ALTAI, p. 13; see also AEPD Audit requirements, UK ICO Guidance and CNIL Self-assessment guide for AI systems)			P	
11.47) Did you consider the privacy and data protection implications of the AI system's non-personal training-data or other processed non-personal data? (ALTAI p. 13; see also AEPD Audit requirements, UK ICO Guidance and CNIL Self-assessment guide for AI systems)			P	
11.48) Did you align the AI system with relevant standards or widely adopted protocols for (daily) data management and governance? (ALTAI p. 13; CNIL Self-assessment guide for AI systems: Achieving compliance; AEPD Audit requirements p. 29)			P	
11.49) Where the AI system / AI-based component is used for in the framework of processing personal data, did you document the intended purpose of the component both in qualitative and quantitative terms? (AEPD Audit requirements p. 15)		Document each purpose for using personal data at each stage of the AI lifecycle, assess whether they are compatible with the originally defined purpose, and schedule reviews to reassess your purposes and whether they remain compatible.	P	
11.50) In the same condition as above, did you also document the relationship between the goal pursued by the use of the AI component in a given processing operation and the conditions guaranteeing the lawfulness of such processing? (AEPD Audit requirements p. 15; see also UK ICO Guidance 29-37; CNIL Self-assessment guide for AI systems: Asking the right questions before using an artificial intelligence system)			P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
12) Accessibility and universal design				
12.1) Did you ensure that the AI system accommodates a wide range of individual preferences and abilities? (ALTAI p. 17)		The requirement elicitation phase will consider a wide range of actual user and stakeholder needs	P	
12.1.1) Did you assess whether the AI system and its user interface is usable by those with special needs or disabilities or those at risk of exclusion? How was this designed into the system and how is it verified? (ALTAI, p. 17)		The verification task will consider a usability analysis at design and testing phases. In the testing phase the usability analysis involve users with special needs or disabilities or those at risk of exclusion	P	
12.1.2) Did you ensure that information about the AI system is accessible and usable also to users of assistive technologies? (ALTAI, p. 17)		The verification task will include an accessibility analysis of AI system also to users of assistive technologies	P	
12.1.3) Did you involve or consult this community during the planning and development phase of the AI system? (ALTAI, p. 17)		Potential users and stakeholder will be involved during the entire project in the requirement, dissemination and ethical assessment tasks	P	
12.2) Did you take the impact of your AI system on the potential user audience into account? (ALTAI, p. 17)		An impact assessment will be carried out both in the proposal preparation and detailed in the first phase of the project. It will take the impact of the AI system on the potential user audience	P	
12.2.1) Did you assess whether the team involved in building the AI system is representative of and engaged with your target user audience? Is it representative of the wider population, considering also of other groups who might tangentially be impacted? (ALTAI, p. 17)			P	
12.2.2) Did you assess whether there could be persons or groups who might be disproportionately affected by negative implications? (ALTAI, p. 20)		As part of the impact assessment activities, assess whether there could be persons or groups who might be disproportionately affected by negative implications	P	
12.4) Did you assess the risk of the possible unfairness of the system onto the end-user's or subject's communities? (ALTAI, p. 17; UK ICO Guidance, p. 104 and following)			P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
13) Stakeholder participation				
13.1) Did you consider a mechanism to include the participation of different stakeholders in the AI system's development and use? (ALTAI p. 18; see also UK ICO guidance, p. 109-110; CNIL Self-assessment guide for AI systems: Developing and training an algorithm; AEPD Audit requirements p. 14)		Potential users and stakeholder will be involved during the entire project in the requirement, dissemination and ethical assessment tasks Option 1) include a knowledge exchange forum to share with the stakeholder information but also co-design aspects	P	
13.2) Did you pave the way for the introduction of the AI system in your organisation by informing and involving impacted workers and their representatives in advance? (ALTAI, p. 20)			P	
14) Sustainable and environmentally friendly AI				
14.1) Did you consider potential negative impacts of the AI system on the environment? Did you establish mechanisms to measure the environmental impact of the AI system's development, deployment and use (for example the type of energy used by the data centres)? (ALTAI, p. 19)		As part of the Impact assessment, measure the environmental impact of the AI system's development, deployment and use (for example the type of energy used by the data centres)	P	
14.2) Did you ensure measures to reduce the environmental impact of your AI system's life cycle? (ALTAI, p. 19)		In the design phase of the AI system identify measures to reduce the environmental impact of AI system's life cycle	T	
15) Social impact				
15.1) In case the AI system interacts directly with humans:				
15.1.1) Did you assess whether the AI system encourages humans to develop attachment and empathy towards the system? (ALTAI, p. 8)		In case the AI system interacts directly with humans, include in the impact assessment an evaluation whether the AI system encourages humans to develop attachment and empathy towards the system	P	
15.1.2) Does the AI system simulate social interaction with or between end-users or subjects? Did you ensure that the AI system clearly signals that its social interaction is simulated and that it has no capacities of "understanding" and "feeling"? (ALTAI p. 7 and 15)		In the design phase of the AI system develop a mechanism that signals that AI system social interaction is simulated and that it has no capacities of "understanding" and "feeling" (In case the AI	P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
		system interacts directly with humans)		
15.2) Did you ensure that the social impacts of the AI system are well understood? For example, did you assess whether there is a risk of job loss or de-skilling of the workforce? What steps have been taken to counteract such risks? (ALTAI, p. 20)		As part of the impact assessment activities, ensure that the social impacts of the AI system are well understood. For example, assess whether there is a risk of job loss or de-skilling of the workforce	P	
15.3) Does the system promote or require new (digital) skills? Did you provide training opportunities and materials for re- and up-skilling? (ALTAI, p. 20)			P	
16) Society and democracy				
16.1) Did you assess the broader societal impact of the AI system's use beyond the individual (end-)user, such as potentially indirectly affected stakeholders or society at large? (ALTAI, p. 20)		As part of the impact assessment activities, assess the broader societal impact of the AI system's use beyond the individual (end-)user, such as potentially indirectly affected stakeholders		
16.2) Did you take action to minimize potential societal harm of the AI system? (ALTAI, p. 20)			P	
16.3) Did you take measures that ensure that the AI system does not negatively impact democracy? (ALTAI, p. 20)			P	
16.4) Have the consequences for groups of individuals based on gender, origin, political or religious opinions, for society and democracy in general, been considered? (CNIL Self-assessment guide for AI systems: Ensuring individuals can fully exercise their rights)			P	
Reliable and function as intended , minimizing unintentional and unexpected harm, preventing unacceptable harm and safeguarding the physical and mental integrity of humans				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI)</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall project, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
17) Fallback plan and general safety				
17.1) Did you ensure that your system has a sufficient fallback plan if it encounters adversarial attacks or other unexpected situations (for example technical switching procedures or asking for a human operator before proceeding)? (ALTAI, p. 11)		In the design phase of the AI system ensure that the system has a sufficient fallback plan if it encounters adversarial attacks or other unexpected situations (for example technical switching procedures or asking for a human operator before proceeding)	T	
17.2) Did you consider the risks, risk metrics and level of risks raised by the AI system in a specific use case? (ALTAI, p. 10)		As part of the impact assessment activities, consider the level of risk raised by the AI system in several socio-technical scenario	P	
17.2.1) Did you put any process in place to measure and assess risks and safety? (ALTAI, p. 10)		As part of impact assessment activities, put in place a process to measure and assess risks and safety	P	
17.2.2) Did you provide the necessary information in case of a risk for human physical integrity? (ALTAI, p. 10-12)		In case of a risk for human physical integrity, you will provide all the necessary information to prevent, detect and mitigate such risks	P	
17.2.3) Did you consider an insurance policy to deal with potential damage from the AI system?		Consider an insurance policy to deal with potential damage from the AI system	P	
17.2.4) Did you identify potential safety risks of (other) foreseeable uses of the technology, including accidental, inappropriate or malicious use or misuse? Is there a plan to mitigate or manage these risks? Do you provide instructions concerning this aspect? (ALTAI, p. 10)		Carry out a safety risk assessment to identify potential safety risks of (other) foreseeable uses of the technology, including accidental or malicious misuse. As part of the safety risk assessment define a plan to mitigate or manage these risks.	P	
17.2.5) Did you define safety criticality levels of the possible consequences of faults or misuse of the AI system? (ALTAI, p. 10)			T	
17.3) Did you assess whether there is a probable chance that the AI system may cause damage or harm to users or third parties? Did you assess the likelihood, potential damage, impacted audience and severity? (ALTAI, p. 10, 15)		As part of the safety risk assessment activities, assess whether there is a probable chance that the AI system may cause damage or harm to users or	P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
		third parties. Moreover assess the likelihood, potential damage, impacted audience and severity		
17.3.1) Did you consider the liability and consumer protection rules, and take them into account? (ALTAI p.29)		Consider and take into account the liability and consumer protection rules	P	
17.3.2) Did you consider the potential impact or safety risk to the environment or to animals? (ALTAI p.29)		As part of the safety risk assessment activities, consider the potential impact or safety risk to the environment or to animals	P	
17.3.3) Did your risk analysis include whether security or network problems such as cybersecurity hazards could pose safety risks or damage due to unintentional behaviour of the AI system? (ALTAI p.29)		As part of the safety risk assessment activities, analyse whether security or network problems such as cybersecurity hazards could pose safety risks or damage due to unintentional behaviour of the AI system	P	
17.4) Did you estimate the likely impact of a failure of your AI system when it provides wrong results, becomes unavailable, or provides societally unacceptable results (for example discrimination)? (ALTAI, p.10-11, see also UK ICO Guidance, UK ICO AI and data protection toolkit, AEPD Audit requirements p. 16)		As part of the safety verification activities, carried out at design and testing phases, estimate the likely impact of a failure of your AI system when it provides wrong results, becomes unavailable, or provides societally unacceptable results (for example discrimination)	P	
17.4.1) Did you assess the dependency of a critical AI system's decisions on its stable and reliable behaviour? (ALTAI, p. 10)			T	
17.4.2) Did you define thresholds and did you put governance procedures in place to trigger alternative/fallback plans? (ALTAI, p. 10)		As part of the safety risk assessment, define thresholds and put governance procedures in place to trigger alternative/fallback plans	P	
17.4.3) Did you define and test fallback plans? Did you establish a post-monitoring plan for the AI system? (ALTAI, p. 10)		As part of the safety risk assessment and verification activities, fallback plans will be defined (risk assessment) and tested (verification at testing phase)	P	
17.5) Did you develop a mechanism to evaluate when the AI system has been changed to merit a new review of its technical robustness and safety? (ALTAI, p. 10)			T	
17.6) Did you consider appropriate quality management system and conformity assessment			T and P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
procedure before providing the AI system? (ALTAI, p. 14; see also CNIL Self-assessment guide for AI systems:Developing and training an algorithm)				
17.7) Have you provided and prepare technical documentation of the AI system before providing it for the use and did you review it and keep updated when using it? (AI Act, Art. 11; ALTAI p. 11; see also UK ICO AI and data protection toolkit; AEPD Audit requirements p. 13)			T and P	
18) Ability to redress				
18.1) Did you establish an adequate set of mechanisms that allows for redress in case of the occurrence of any harm or adverse impact? (ALTAI, p. 22)		In the design phase of the AI system, define a set of mechanisms that allows for redress in case of the occurrence of any harm or adverse impact. These mechanism will be tested in the verification phase of the project	T	
18.2) Did you put mechanisms in place both to provide information to (end)users/third parties about opportunities for redress? (ALTAI, p. 22)		In the design phase of AI system, put in place mechanisms to provide information to (end-)users/third parties about opportunities for redress	T	
Able to provide a suitable explanation of its decision making process , whenever an AI based system can have a significant impact on people's lives				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall project, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
19) Explainability				
19.1) Did you assess:				
19.1.1) to what extent the decisions and hence the outcome made by the AI system can be understood by different stakeholders? (ALTAI p. 31)		As part of the Impact Assessment, assess to what extent the decisions and hence the outcome made by the AI system can be understood to decide how comprehensive the explanation needs to be.	P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
19.1.2) to what degree the system's decision influences the organisation's decision-making processes? (ALTAI p.31)		As part of the Impact Assessment, assess to what degree the system's decision influences the organisation's decision-making processes	P	
19.1.3) why this particular system was deployed in this specific area? (ALTAI, p. 31)		As part of the Impact Assessment, assess why this particular system was deployed in this specific area	P	
19.1.4) are the elements used for decision-making by the AI system logged? (see AI Act, Art. 12; ALTAI, p. 21; UK ICO AI and data protection toolkit)			T	
19.1.5) what the system's business model is (for example, how does it create value for the organisation)? (ALTAI, p. 31)		As part of the exploitation task, define business model of the AI system	P	
19.1.6) the explainability and transparency requirements considering the domain, sector or use case that the AI system will be deployed in? (see ALTAI, p. 14-15; UK ICO Guidance p. 48; UK ICO AI and data protection toolkit)		Assess people's expectations of the content and scope of similar explanations previously offered or researching sector-specific expectations. For example, what are the expectations of individuals in a health context compared to an insurance context. Consider, for example, the people affected by the decision and the end users.	P	
19.2) Did you ensure an explanation as to why the system took a certain choice resulting in a certain outcome that all users can understand? (ALTAI, p. 14; see also UK ICO AI and data protection toolkit. UK ICO Guidance p. 94, 97)		As part of the verification activities made at design and testing phases, ensure an explanation as to why the system took a certain choice resulting in a certain outcome that all users can understand. Test the effectiveness of your explanations by measuring how well individuals can understand why the model made the decision it did, or how the model output contributed to the decision. (UK ICO AI and data protection toolkit)	P	
19.3) Did you design the AI system with interpretability in mind from the start? (see UK ICO guidance, p. 96, 100, 102, AEPD Audit requirements, p. 23)		Design the AI system taking into account the interpretability principle	T	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
19.3.1) Did you define the levels / thresholds for interpreting and using the results yielded by the relevant AI-based component / model? (AEPD Audit requirements, p. 16)			T	
19.3.2) Did you research and try to use the simplest and most interpretable, appropriate model possible for the application in question and for training the AI-based component? (in consideration of parameters such as efficiency, quality, accuracy, explainability, logical comprehension, readability, and internal consistency) Did you provide this information in a transparent and clear and simple terms to the stakeholders? (AEPD Audit requirements, p. 14)		In the design phase of the AI system, research and try to use the simplest and most interpretable model possible for the application in question	T	
19.3.3) Did you assess whether you can analyse your training and testing data? Can you change and update this over time? (see UK ICO Guidance, UK ICO AI and data protection toolkit, AEPD Audit requirements, p. 20)		As part of the Impact Assessment activities, assess whether training and testing data can be analysed, changed or updated, e.g. it may be needed to go back into the training stage and redefine the target variable or change the labels in the input data.	P	
19.3.4) Did you assess whether you can examine interpretability after the model's training and development, or whether you have access to the internal workflow of the model? (ALTAI, p31)		As part of the verification activities made at design and testing phases examine interpretability after the model's training and development and whether the internal workflow of the model is accessible	P	
19.3.5) Did you assess and document explainability of your AI system and consider what supplementary tools you can use to help explain decisions made by your AI system to individuals who will be affected? (UK ICO AI and data protection toolkit)		Assessment and documentation of explainability are required to prevent an unexplainable model being deployed and negatively impacting a population. Provision of meaningful information about the logic involved where solely automated decision-making is used which has legal or similarly significant effects for the individual. Design explanations that meet the needs of those interacting with the system at different moments are required. Consider textual	T and P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
		clarification, visualisation media, graphical representations, summary tables, or a combination as part of explanation. (UK ICO AI and data protection toolkit)		
19.4) Can the logged elements be used to explain, retrospectively, a particular decision made by the AI system? (CNIL Self-assessment guide for AI systems: Using an AI system in production; ALTAI, p. 14) CNIL Guidelines			T	
19.5) Did you define feedback mechanism that allows individuals to comment on the explanations they receive? (UK ICO AI and data protection toolkit)		A feedback mechanism would allow for better detection of ineffective explanations and their correction. (UK ICO AI and data protection toolkit)	T	
20) Communication				
20.1) Did you communicate to (end-)users that they are interacting with an AI system and not with another human? Did you label your AI system as such? (ALTAI, p. 15, CNIL Self-assessment guide for AI systems: Using and AI system in production)		In the design phase of the AI define a mechanism to communicate to (end-)users – through a disclaimer or any other means – that they are interacting with an AI system and not with another human	T	
20.2) Did you establish mechanisms to inform (end-)users on the reasons and criteria behind the AI system's outcomes? (ALTAI, p. 15)		In the design phase of the AI define mechanisms to inform end-)users on the reasons and criteria behind the AI system's outcomes	T	
20.2.1) Did you communicate this clearly and intelligibly to the intended audience? (ALTAI, p. 15; CNIL Self-assessment guide for AI systems: Using and AI system in production)			P	
20.2.2) Did you establish processes that consider users' feedback and use this to adapt the system? (ALTAI, p. 18; see also AEPD Audit requirements,)		As part of the Ethical processes, establish a process that consider users' feedback and use this to adapt the system	P	
20.2.3) Did you communicate potential or perceived risks, such as bias, as well as technical limitations, such as the level of accuracy and/or error rates? (ALTAI, p. 15)		In the design phase of the AI system, define a mechanism to communicate around potential or perceived risks, such as bias	T	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
20.2.4) Did you consider communication and transparency towards other audiences, third parties or the general public?				
20.3) Did you clarify the purpose of the AI system and who or what may benefit from the product/service? (ALTAI, p. 15)		As part of the Impact Assessment, clarify the purpose of the AI system and who or what may benefit from the product/service	P	
20.4) Did you specify usage scenarios for the product and clearly communicate these to ensure that it is understandable and appropriate for the intended audience/users? (ALTAI, p.31)		As part of the Impact assessment specify usage scenarios for the product. Use the analysed scenario in the awareness and training task.	P	
20.5) Depending on the use case, did you think about human psychology and potential limitations, such as risk of confusion, confirmation bias or cognitive fatigue? (ALTAI p31)		As part of the Impact Assessment, consider human psychology and potential limitations, such as risk of confusion, confirmation bias or cognitive fatigue	P	
20.6) Did you clearly communicate characteristics, limitations, capabilities and potential shortcomings of the AI system? (ALTAI, p. 15)		As part of the verification activities analyze limitation and potential shortcomings of the AI system	P	
20.6.1) In case of the system's development: to whoever is deploying it into a product or service?			P	
20.6.2) In case of the system's deployment: to the (end-)user or consumer?			P	
20.7) Did you provide instructions in digital format concerning the development, training and deployment of the AI system? (AI Act, Art. 13)			P	
21) Traceability				
21.1) Did you establish measures that can ensure traceability of the AI system in its whole life cycle? (ALTAI, p. 14) This could entail documenting the following methods:		In the design phase of the AI system, establish measures that can ensure traceability.	T	
21.1.1) Methods used for designing and developing the algorithmic system: - Rule-based AI systems: the method of programming or how the model was built; - Learning-based AI systems: the method of training the algorithm, including which input data was gathered and selected, and how this occurred. (ALTAI, p.30)		The design documentation should contain the methods used for designing and developing the algorithmic system	P	
21.1.2) Methods used to test and validate the algorithmic system: Rule-based AI systems; the scenarios or cases used in order to test and validate;		The validation documentation will contain the methods used to test	P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
Learning-based model: information about the data used to test and validate. (ALTAI, p.31)		and validate the algorithmic system		
21.1.3) Outcomes of the algorithmic system: The outcomes of or decisions taken by the algorithm, as well as potential other decisions that would result from different cases (for example, for other subgroups of users). (ALTAI, p. 31)		The operational documentation will contain the outcomes of or decisions taken by the algorithm, as well as potential other decisions that would result from different cases (for example, for other subgroups of users).	P	
21.2) Did you document the following information? Identifying criteria of the AI-based component (i.e. code, version of the component, date of creation); Version history of the AI-based component, including information on the parameters used to train the component. (AEPD Audit requirements p. 13)			P	
21.3) Is there a control version system in place for all elements of the AI-based component: used datasets, AI-based component code, libraries used and any other element associated with the component? (AEPD Audit requirements p. 28)			T	
21.4) Did you implement appropriate monitoring, verification and supervision mechanisms (e.g. log files and results records) for the AI-based component, enabling the assessment of component's behavior while interacting with its environment? (AEPD Audit requirements p. 28; see also UK ICO Guidance)			T and P	
21.5) Did you detect and document any previous incidences and abnormal behaviors? (AEPD Audit requirements p. 29; see also UK ICO Guidance)			P	
21.6) Did you put in place measures to continuously assess the quality of the input and output data to the AI system? (ALTAI, p. 14; see also CNIL Self-assessment guide for AI systems and UK ICO guidance)			P	
21.7) Did you establish procedures for human intervention in instances where the output of the component diverges from what is expected as well as when data subjects affected by the outputs of the component request so? (AEPD Audit requirements, p. 29; see also CNIL Self-assessment guide for AI systems; UK ICO Guidance; ALTAI)			P	
21.8) Can you trace back which data was used by the AI system to make a certain decision(s) or recommendation(s)? (ALTAI, p. 14)			T and P	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
21.9) Can you trace back which AI model or rules led to the decision(s) or recommendation(s) of the AI system? (CNIL Self-assessment guide for AI systems: Using an AI system in production)			T and P	
21.10) Are the learning and production phases of the AI system separate? (CNIL Self-assessment guide for AI systems: Asking the right questions before using an artificial intelligence System)		Part of the assessment of the risks associated with the proposed system and its proportionality ahead of the design stage	T and P	
21.10.1) If yes to the above question, is a second assessment planned for the production phase? (CNIL Self-assessment guide for AI systems: Asking the right questions before using an artificial intelligence System)		Part of the assessment of the risks associated with the proposed system and its proportionality ahead of the design stage	T and P	
22) Accountability				
22.1) If personal data is collected and/or used, has a data controller and data processor(s) been identified? (CNIL Self-assessment guide for AI systems: Asking the right questions before using an artificial intelligence System; see also UK ICO Guidance; AEPD Audit requirements)			P	
22.2) Are the legal and natural persons in charge of the AI system development, deployment, maintenance, error corrections and monitoring clearly defined? (CNIL Self-assessment guide for AI systems: Asking the right questions before using an artificial intelligence System; see also UK ICO Guidance; AEPD Audit requirements)			P	
22.3) Have they been made aware of the technical and moral issues of AI? (CNIL Self-assessment guide for AI systems: Asking the right questions before using an artificial intelligence System)			P	
22.4) Is there a policy governing the design and deployment of AI systems? (CNIL Self-assessment guide for AI systems: Asking the right questions before using an artificial intelligence System)			P	
22.5) Is the development of the AI-based component documented in a manner that makes the aspects related to its implementation, operating, and interconnection with other processing elements understandable? (AEPD Audit requirements p. 20-21)			P	
22.6) Is the basic architecture of the AI-based component identified and documented (incl. information on the chosen machine learning technique, types of tested and dismissed algorithms for the learning and training of the AI			P and T	

Technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk posed by the AI based system or technique				
<u>Questions from Assessment list of Trustworthy Artificial Intelligence (ALTAI) (see Assessment List of Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission) and supplementary questions from UK ICO Guidelines (UK ICO Guidance and UK ICO AI and data protection toolkit), CNIL Self-assessment guide for AI systems and AEPD Audit Requirements for Personal Data Processing Activities involving AI</u>	<u>Answers to the Assessment List (Yes/No)</u>	<u>Possible measures to consider in the overall application of the AI based system, or in the design and development or in the use of AI based System</u>	<u>Technical (T) or Procedural (P) measures</u>	<u>Development (D) or Use (U)</u>
component, any other data on the functioning and implementation of the component - i.e. what the component does, why the component works in a certain way, what quality and legibility is the code of)? (AEPD Audit requirements p. 20-21)				
22.7) Did you document and implement strict controls over who has access to training data, training code, and deployment code? (UK ICO AI and data protection toolkit)		It is relevant to ensure there is a clear audit trail of who has access to the training data, training code, and deployment code, and when they have access.	T	
23) Auditability				
23.1) Did you establish mechanisms that facilitate the system's auditability, such as ensuring traceability and logging of the AI system's processes and outcomes? (ALTAI, p. 21)		In the design phase of the AI system, establish mechanisms that facilitate the system's auditability, such as ensuring traceability and logging of the AI system's processes and outcomes	T	
23.2) Did you ensure that the AI system can be, where necessary, audited by independent third parties? Did you ensure, in applications affecting fundamental rights (including safety-critical applications) that the AI system can be audited independently? (ALTAI, p. 21)		As part of the verification activities made at design and testing phases, ensure, in applications affecting fundamental rights (including safety-critical applications) that the AI system can be audited independently	P	
23.3) Did you define the methods to keep samples of training data used for the training of the AI-based component for the purposes of auditing the component, in a manner where the method poses the least amount of risks to data subjects? (AEPD Audit requirements, p. 17-18; UK ICO AI and data protection toolkit)			T and P	

15. ANNEX 5: GDPR COMPLIANCE CHECKLIST FOR TAI-SDF AND STM

GDPR compliance questionnaire for Data Providers and Data Consumers

The list is based on the provisions of the GDPR and is a translation of additional requirements indicated in D2.1. for Data Providers and Consumers in the context of sharing and processing of personal data.

No	Questions based on the GDPR and selected LEG-USR requirements included in D2.1, pertaining to data protection aspects	Answers to the questions (Yes / No / N/A)
1	Is the data you process (encrypted) personal data?	Yes – further questions apply No – only Q36-38 apply
2	For Data Providers – do you share your data in anonymised form?	Yes – accepted and no other questions applicable No – all questions in the questionnaire applicable
3	For Data Providers – do you share your data in pseudonymised form?	Yes – accepted No – justification why not pseudonymised required, but the lack of pseudonymisation does not mean lack of compliance N/A
4	Do you consider and adhere to the personal data processing principles as defined in Article 5 GDPR? <i>The principles which have to be considered are:</i> - Lawfulness, fairness and transparency - Purpose limitation - Data minimisation - Accuracy - Storage Limitation - Integrity and confidentiality - Accountability	Yes – accepted No – risk of lack of compliance
5	Do you have a legal basis/legal ground for: - for Data Providers - sharing the data with the Consumers / making the data available to Consumers via TRUSTEE? - for Consumers - processing the data obtained from Data Providers? <i>Consumer should indicate legal basis for processing of personal, HE encrypted data, provided by Data Provider, and, where applicable, indicate exemption from Article 9 for processing of</i>	Yes – accepted No – risk of lack of compliance

	<i>special categories of data (such as data concerning health, genetic data).</i>	
6	Did you consider if specific national laws or ethical standards require additional consent and permissions or a specific legal basis for: - for Data Providers - sharing the data with Consumer? - for Consumers – for processing the data obtained from Data Provider?	Yes – accepted – <i>it would be good if a person could provide specific information about applicable laws</i> No – risk of lack of compliance
6.1	Do you adhere to those standards?	Yes – accepted No – risk of lack of compliance
7	Do you process your data jointly with Data Provider/ Consumer?	Yes – answer to question 7.1 required to accept the point No – accepted, skip question 7.1 and move to 8
7.1	If yes to the above question, did you establish and conclude a Joint Controllership Agreement for joint processing of the data with Data Provider/Consumer for joint purposes and with joint means?	To be answered only if Q7 is YES Yes – accepted No – risk of lack of compliance
8	Do you involve any data processor to process the data on your behalf?	Yes – answer to Q8.1 required to accept the point No – accepted, skip Q8.1 and move to 9
8.1	Did you conclude a Data Processing Agreement with the data processor that includes the requirements listed by Article 28 of the GDPR?	Yes – accepted No – risk of lack of compliance
9	For Data Providers - did you provide data subjects with the information that their data will be processed / shared via TRUSTEE and share categories of recipients of data (i.e. categories of Consumers), according to Article 13 GDPR?	Yes – accepted No – risk of lack of compliance – additional justification required
10	For Consumers - if possible for you, did you inform data subjects about the purposes their data will be processed for and the manner how they will be processed, (according to Article 14 GDPR)?	Yes – accepted No – additional justification based on Article 14(5) required, if not justified – risk of lack of compliance
11	For Data Providers – did you indicate for what purposes data can be processed by the Consumers?	Yes – accepted

		No – risk of lack of compliance
12	Did you sign a data sharing agreement including the following information? a) the parties identification and information on the participants of data sharing, b) the purpose of data sharing defined in a clear manner, i.e. description of the aims for which data are shared, why sharing is necessary to achieve those goals and what benefits to individuals and/or to society the parties hope to bring, c) what data are going to be shared d) lawful basis for data sharing (both parties need to clarify it for themselves – Data providers the basis for making data available, and Consumers for accessing the data) e) information if there is any type of special categories of data that would be shared f) procedures allowing for achieving compliance with the rights of individuals (data subjects) g) governance arrangements between the parties	Yes – accepted No – risk of lack of compliance
13	For Consumers – did you provide the information concerning processing the data (whenever the Consumer processes personal (encrypted) data provided by the Data Provider) and check whether it will correspond with the purposes indicated by Data Provider and whether it is allowed to process the data in a foreseen manner?	Yes – accepted No – risk of lack of compliance
14	For Data Providers – did you conduct compatibility test with the initial purpose and check if you can share the data for secondary use via TRUSTEE? <i>In case the purposes are not compatible, Data Provider should abstain from processing the data or provide another lawful basis and inform data subjects about new processing, where possible. The outcome of the compatibility assessment should be indicated by Data Provider in STM and in the context of the DPIA.</i>	Yes – accepted No – risk of lack of compliance
14.1	Is your processing foreseen for archiving, historical or scientific research purposes based on the EU or national law and considered compatible with the initial purpose you collected and processed the data for? <i>In that case purpose is considered compatible by default, and appropriate technical and organisational safeguards (measures) must be applied.</i>	Yes – accepted No - accepted

15	For Data Providers – did you disclose the results of the compatibility assessment in the data sharing agreement concluded with the Consumer?	Yes – accepted No – further explanation required
16	Do you process only the data that are necessary to achieve indicated purpose? <i>Specification should be included in the terms of use of the data (STM) and in the DPIA, and be lawful.</i>	Yes – accepted No – risk of lack of compliance
17	For Data Providers - did you ensure that the data provided via TRUSTEE are accurate and of appropriate quality? Did you indicate that in the data sharing agreement?	Yes – accepted No – risk of lack of compliance
18	For Data Providers - when a data subject exercises its right of erasure and/or restriction to the processing, do you ensure to remove the data from TRUSTEE / stop sharing the data of data subjects when requested by a data subject?	Yes – accepted No – justification required, otherwise risk of lack of compliance
18.1	For Data Providers – do you ensure to remove the data from TRUSTEE / stop sharing the data of data subjects when required by the data protection laws?	Yes – accepted No – risk of lack of compliance
18.2	For Data Providers - when a data subject exercises its right of erasure and/or restriction to the processing, do you ensure to remove the data from TRUSTEE / stop sharing the data of data subjects when the data subject withdraws consent and there is no legal basis for further sharing and processing of the data?	Yes – accepted No – risk of lack of compliance
19	Did you set storage limits (how long the data will be stored)?	Yes – accepted No – risk of lack of compliance
20	Do you apply the means to allow the data subjects to exercise their rights to: • Information about the data processing (Article 12-14) • Access (Article 15 GDPR) • Rectification (Article 16 GDPR) • Erasure (Article 17 GDPR) • Restriction on processing (Article 18 GDPR) • Portability (Article 20 GDPR) • Object (Article 21 GDPR) • Not to be subject to a decision based solely on automated processing, including profiling (Article 22)	Yes – accepted No – risk of lack of compliance

	GDPR) • Lodge a complaint with the supervisory authority (Article 13 GDPR)?	
21	Does the application of AI models offered by TRUSTEE lead to any automated individual decision-making about a data subject whose data are subject to processing?	Yes – answer Q21.1, Q21.2 and Q21.3 to assess if you are compliant No – accepted, skip Q21.1, Q21.2 and Q21.3
21.1	Is the data subject (whose homomorphically encrypted data is accessed and processed) informed about automated individual decision-making?	Yes – accepted No – risk of lack of compliance, unless justification why it was not possible, in line with Article 13 and 14(5) GDPR, is provided
21.2	Did you provide the information which exception, which allows the automated individual decision making (necessity to perform a contract, authorisation by the EU or national law, explicit consent of a data subject), applies?	Yes – accepted No – risk of lack of compliance
21.3	For special categories of data, did you provide the information whether automated individual decision-making is allowed due to an explicit consent of a data subject or due to a substantial public interest, based on the EU or national law?	Yes – accepted No – risk of lack of compliance
22	Do you apply appropriate technical and organisational measures to protect personal data during processing?	Yes – accepted No – risk of lack of compliance
23	Did you carry out an objective quantitative and qualitative risk assessment of the processing of the data you foresee to conduct? The more severe a potential risk, the higher level of security to mitigate the risk should be.	Yes – accepted No – risk of lack of compliance
24	Do you ensure that the data are authentic and accurate, and have not been maliciously or accidentally altered at any stage (when processed, stored or transmitted)?	Yes – accepted No – risk of lack of compliance
25	Do you access only the data you need?	Yes – accepted No – justification required, otherwise risk of lack of compliance

26	For Consumers – would you try to re-identify a data subject when it is not necessary or forbidden by the Data Provider? <i>Appropriate statements referring to that should be included in the agreements.</i>	Yes – risk of lack of compliance No - accepted
27.1	For Data Providers – did you make available to the Consumer the confidentiality standards which the Consumer should adhere to?	Yes – accepted No – risk of lack of compliance N/A – accepted
27.2	For Consumer – do you adhere to the confidentiality standards required by Data Provider?	Yes – accepted No – risk of lack of compliance N/A - accepted
28	For Data Provider – did you indicate to what countries the data can be made available?	Yes – accepted No – justification required
28.1	For Data Provider – did you indicate that the data can be provided only to the Consumers located in the countries with an adequate level of protection, as deemed by the European Commission?	Yes – accepted No – Q28.2 has to be answered
28.2	For Data Provider – in the case of lack of an adequacy decision for a specific country, did you assess whether the country where the Consumer accessing and processing the data is based ensures appropriate safeguards to protect the data in and after transfer and required the Consumer to provide such safeguards? <i>This should be included in the agreement and assessment report. Those safeguards may be:</i> <i>a. a legally binding and enforceable instrument between public authorities or bodies;</i> <i>b. binding corporate rules in accordance with Article 47 GDPR;</i> <i>c. standard contract clauses adopted by the Commission;</i> <i>d. standard contract clauses adopted by a supervisory authority and approved by the Commission;</i> <i>e. an approved code of conduct together with binding and enforceable commitments to apply safeguards; or</i> <i>f. an approved certification mechanism together with</i>	Yes – accepted No – risk of lack of compliance

	<i>binding and enforceable commitments to apply safeguards.</i>	
28.3	For Data Provider – did you require the Consumer located in non-EU country, and which does not benefit from an adequacy decision by the European Commission, to sign standard contractual clauses / data sharing agreement before obtaining the access to the data?	Yes – accepted No – if answer to Q28.1 or Q28.2 „YES“ – accepted; No – if answer to Q28.1 or Q28.2 „NO“ – justification why required
29	Did you conduct a Data Protection Impact Assessment (DPIA) prior to processing the data if processing is likely resulting in a high risk to the rights and freedoms of the people whose data they process, as well as if two of the following requirements materialise? - if the envisioned processing involves evaluation or scoring, profiling and predicting activities. - if you envisage automated individual decision making with legal or similar significant effects to a person. - if you envisage systematic monitoring. - if you process special categories of data, such as, e.g. health data or genetic data. - if you envisage processing of the data on a large scale. - if you envisage processing of the data of vulnerable individuals - e.g. patients or minors. - if you envisage matching or combining different datasets. - if you envisage data transfer across borders outside the European Union. - if the processing in itself “prevents data subjects from exercising a right or using a service or a contract”.	Yes – accepted No – risk lack of compliance
30	For Data Providers - did you identify risks to data processing and their mitigation and indicate it in the DPIA before granting actual access to (homomorphically encrypted) data?	Yes – accepted No – risk of lack of compliance
31	Did you conduct a DPIA in all the cases where it is not clear/certain to you whether a DPIA is required?	Yes – accepted No – risk of lack of compliance
32	Did you take into account applicable national laws and lists of the activities which require a Data Protection Impact Assessment?	Yes – accepted No – risk of lack of compliance

33	Do you keep the record of the DPIA?	Yes – accepted No – risk of lack of compliance
34	For Data Provider – did you inform Consumer about any applicable internal policies that should be applicable to data sharing and processing?	Yes – accepted No – justification why not
35	Do you keep the record of data processing activities (in line with Article 30 GDPR)?	Yes – accepted No – risk of lack of compliance
36	Did you indicate the limitations for data processing stemming from their intellectual property rights and sui generis protection of databases (contents of databases), computer programmes, software, and AI/ML models (where this qualifies for any IPR protection)? Did you include it in the agreement with Data Provider/Consumer?	Yes – accepted No – risk of lack of compliance
36.1	Did you include the limitations from Q36 in the agreement with Data Provider/Consumer?	Yes – accepted No – risk of lack of compliance
37	For Data Provider – did you give truthful information with respect to intellectual property rights held by them on any type of data (where this qualifies for copyright protection), database, software and/or AI/ML models (where this qualifies for any IPR protection) and make necessary attribution to rightholders, where relevant? In case you are not the rightholder, did you ensure that necessary permissions are obtained before making available any datasets via TRUSTEE?	Yes – accepted No – risk of lack of compliance
38	For Data Providers - did you indicate the limitations and obligations binding for Consumers, stemming from the protection from trade secrets law?	Yes – accepted No – risk of lack of compliance

