SPATIAL

D1.1 REQUIREMENTS ANALYSIS FOR AI TOWARDS ADDRESSING SECURITY RISKS AND THREATS TO SYSTEM AND NETWORK ARCHITECTURES

Revision: v.1.2

Work package	WP1
Task	Task 1.1
Due date	30/04/2022
Submission date	30/05/2023
Deliverable lead	Fraunhofer Institute for Open Communication Systems (FOKUS)
Version	1.2
Authors *(sorted by alphabetical order of partner names)*	João Fernando Ferreira Gonçalves (EUR), Tessa Oomen (EUR), Jorge Pereira Campos (EUR), Michell Boerger (FOKUS), Denis Rangelov (FOKUS), Nikolay Tcholtchev (FOKUS), Samuel Marchal (FSC), Ivan Milosevic (MFX), Vinh Hoa La (MI), Manh Dung Nguyen (MI), Claudio Soriente (NEC), Nicolas Kourtellis (TID), Souneil Park (TID), Prachi Bagave (TUD), Aaron Ding (TUD), Madhusanka Liyanage (UCD), Chamara Sandeepa (UCD), Thulitha Senevirathna (UCD), Bartlomiej Siniarski (UCD), Shen Wang (UCD), Huber Flores (UT), Abdul-Rasheed Ottun (UT)
Reviewers	Ana Cavalli (Montimage) Souneil Park (TID)

Grant Agreement No.: 101021808
Call: H2020-SU-DS-2020
Topic: SU-DS02-2020
Type of Action: RIA

Abstract	This document is part of the first phase of the SPATIAL project, which focuses on identifying requirements and design guidelines for modern system architectures based on accountable AI. The deliverable document at hand is to be regarded as the basis for achieving these goals. Precisely, this document aims to establish a comprehensive catalogue of aspects and general design principles to be considered when integrating and utilizing AI algorithms and frameworks for addressing security risks and threats to system (and network) architectures.
Keywords	AI-based systems, Secure AI, Explainable AI, Accountable AI, Requirements analysis, Trustworthy AI for Cybersecurity, Accountability, Privacy Preservation, Resilience Engineering, Explainable AI

Document Revision History

Version	Date	Description of change	List of contributor(s)
v0.1	01/11/2021	Initial Table of Contents	Nikolay Tcholtchev Michell Boerger
v0.2	21/03/2022	Collected contributions from all involved partners.	All listed authors.
v0.3	31/03/2022	Prepared document for internal review.	Nikolay Tcholtchev Michell Boerger
v0.4	05/04/2022	Integrated contributions of MFX for Section 3.4.1.1.	Nikolay Tcholtchev Michell Boerger
v1.0	25/04/2022	Integrated feedback from internal reviewers.	Nikolay Tcholtchev Michell Boerger
v1.1	09/05/23	Revised document according to review feedback: - Added Section 1.2 to elaborate on the lifecycle of the requirements and the connection of the WP1 activities to other SPATIAL activities (see p.17) - Clarified the process of gathering the requirements and identifying relevant stakeholders in Section 3.1.1 (see p.35) - Provided more details on the prioritization of the requirements in Section 3.1.2 (see p. 38) - Determined deadline for the final set of requirements in the Executive Summary as well as Section 1, Section 3.1.1, and Section 5. - Defined relevance scores for the identified requirements with respect to the other technical activities in WP3 and WP5 in Section 3.1.3 (see, p. 39) - Used these definitions to assign relevance to the requirements in Appendix A (see p.88).	Nikolay Tcholtchev Michell Boerger

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

		- Provided information about the tracking of the requirements in Section 3.1.3 (see p. 39). - Added Section 4 to elaborate on how the requirements can be used to derive design goals and expected functionality of the SPATIAL explanatory AI platform (see p. 74). - Swap Appendix A and B as the requirements table are more important (see p.88 and p.1)	
v1.2	30/05/23	Revised document according to additional feedback: - Sections 1.1 and 1.2: Clarified the relationship between the terms “requirement”, recommendation, and guideline (see p.16, p17 and footnote 1). - Section 3.1.1: Explained the distinction between functional and non-functional requirements and referred to the corresponding mapping provided in the updated requirements tables provided in Appendix A. Furthermore, we added two paragraphs describing how the identified requirements can be used to shape the technical activities in the four SPATIAL use cases and the Explanatory AI platform. In this context, we have referred to the added column “Implemented By” in Appendix A, which illustrates the relationship between the technical activities and the requirements. (see p.35 and p.36) - Section 3.1.3: Adjusted and clarified text to avoid confusion for the relationship between the terms “requirement”, “recommendation”, and “guideline”. - Section 4: Extended Table 1 according to the new structure. Appendix A: Added columns “Functional/Non-functional” and “Implemented By”. The former provides information on whether a listed requirement is considered functional or non-functional. The latter illustrates the relationship between the technical activities and the gathered requirements (see p. 88)	Nikolay Tcholtchev Michell Boerger

DISCLAIMER

The information, documentation and figures available in this deliverable are written by the SPATIAL project’s consortium under EC grant agreement 101021808 and do not necessarily reflect the views of the European Commission.

SPATIAL project is funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 101021808.

The European Commission is not liable for any use that may be made of the information contained herein.

COPYRIGHT NOTICE

Project funded by the European Commission in the H2020 Programme		
Nature of the deliverable:	**R**	
Dissemination Level		
PU	Public, fully open, e.g., web	✓
CL	Classified, information as referred to in Commission Decision 2001/844/EC	
CO	Confidential to SPATIAL project and Commission Services	

* R: Document, report (excluding the periodic and final reports)

DEM: Demonstrator, pilot, prototype, plan designs

DEC: Websites, patents filing, press & media actions, videos, etc.

OTHER: Software, technical diagram, etc

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

EXECUTIVE SUMMARY

This document is part of the first phase of the SPATIAL project, which focuses on identifying requirements and design guidelines for modern system architectures based on accountable Artificial Intelligence (AI). The current document is to be regarded as the basis for achieving these goals. Precisely, this document aims to establish a comprehensive catalogue of aspects and general design principles to be considered when integrating and utilizing AI algorithms and frameworks for addressing security risks and threats to system (and network) architectures.

On the one hand the requirements are identified based on four use cases of industrial and social relevance: (1) the utilization of privacy preserving AI in the cloud-fog-edge continuum, (2) improving the explainability, resilience and performance of cybersecurity in 4G/5G/6G and Internet of Things (IoT) networks, (3) the utilization of accountable AI in next generation emergency communication and (4) resilient cybersecurity analysis based on machine learning models. These use cases - besides an extensive literature review - serve as the basis for cataloguing a first set of requirements for the emerging SPATIAL tools and frameworks for privacy preserving and accountable AI in various highly distributed system architectures. The four use cases and the belonging security and threat analysis are one of pillars for extracting and listing specific needs of modern system architectures in relation to the application of ML models for cybersecurity. In the course of the requirements analysis, we discuss and define the specific stakeholders - such as end users, developers, testers, system operators and others - which are relevant for the emerging SPATIAL eco-system and bring in a special view on the overall framework and tools to emerge in the scope of the project. Finally, all these discussions are used as the foundation for cataloguing specific tangible requirements, which are classified as follows: software and hardware requirements, data requirements, model requirements, legislative requirements, security requirements, usability and finally accessibility requirements.

The above listed contributions provide an initial analysis and set the way forward for the SPATIAL project as a whole. The defined aspects can help to develop more secure, explainable, and trustworthy AI-based systems and security solutions. They aim at providing realistic guidelines for developers and operators on how to design, deploy, and modify AI-based systems, in order to provide streamlined application of secure, more transparent, explainable, and trustworthy AI.

The requirements gathered in this document represent just an initial set of design guidelines and recommendations to consider when integrating and utilizing AI algorithms into modern system architectures. As the project progresses, the requirements and design guidelines will be refined in an agile manner by incorporating the feedback and insights gained from SPATIAL's technical activities of other work packages. Subsequently, an updated and final set of requirements and design guidelines will be provided in deliverable D1.3 "*Final Requirements Analysis for AI towards Addressing Security Risks and Threats to System and Network Architectures*".

TABLE OF CONTENTS

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

SPATIAL project is funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 101021808.

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

LIST OF FIGURES

SPATIAL project is funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 101021808.

LIST OF TABLES

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

ABBREVIATIONS

AI	Artificial Intelligence
API	Abstract Programming Interface
ASIC	Application Specific Integrated Chip
AUC	Area Under the Curve
B5G	Beyond 5G
BAN	Body Area Network
BBU	Baseband Unit
CACE	Changing Anything Changes Everything
CAM	Class Activation Mapping
CNN	Convolutional Neural Network
CPS	Cyber-Physical System
CPU	Central Processing Unit
CRAN	Centralized/Cloud RAN
CRN	Cognitive Radio Networks
D&R	Detection and Response
DBA	Distributed Backdoor Attack
DDoS	Distributed Denial of Service
DHCP	Dynamic Host Configuration Protocol
DL	Deep Learning
DMP	Data Management Plan
DNN	Deep Neural Network
DNS	Domain Name Server
DoS	Denial of Service
DP	Differential Privacy
DPIA	Data Protection Impact Assessment
DPO	Data Protection Officer
DRA	Data Reconstruction Error
EC	European Commission
eCall	Emergency Call
EMYNOS	nExt generation eMergencY commuNicatiOnS

EPC	Evolved Packet Core
ESRP	Emergency Services Routing Proxy
EU	European Union
EUR	Erasmus University Rotterdam
FL	Federated Learning
FLaaS	Federated Learning as a Service
FOKUS	Fraunhofer Institute for Open Communication Systems
FPGA	Field Programmable Gate Array
FSC	F-Secure OYJ
GAP	Global Average Pooling
GDPR	General Data protection Regulation
GPU	Graphics Processing Unit
Grad-CAM	Gradient-weighted Class Activation Mapping
GUI	Graphical User Interface
HAL	Hardware Abstraction Layer
HELD	HTTP-Enabled Location Delivery
HTTP	Hypertext Transfer Protocol
IDS	Intrusion Detection System
IoT	Internet of Things
IP	Internet Protocol
ISP	Internet Service Provider
IT	Information Technology
IVR	Interactive Voice Response
LDAP	Lightweight Directory Access Protocol
LIME	Local Interpretable Model-agnostic Explanations
LIS	Location Information Server
LLDP-MED	Link Layer Discovery Protocol - Media Endpoint Devices
LRP	Layer-Wise Relevance Propagation
LSD	Location to Service Determination
LTE	Long-Term Evolution
MAC	Medium Access Control
MFX	Mainflux Labs
MI	Montimage EURL

MIA	Membership Interference Attack
MITM	Man-in-the-middle
ML	Machine Learning
ML-IDS	Machine Learning based Intrusion Detection System
MLP	Multi-Layer Perceptron
MMT	Montimage Monitoring Tool
NEC	NEC Laboratotories Europe GmbH
NFV	Network Function Virtualization
NG112	Next Generation 112
NIST	National Institute of Standards and Technology
NN	Neural Network
ORAN	Open RAN
PDN	Public Data Network
PDP	Partial Dependence Plot
PIA	Property Interference Attack
PSAP	Public-Safety Answering Point
RAN	Radio Access Network
RCA	Root Cause Analysis
ReLU	Rectified Linear Unit
RFA	Robust Federated Aggregation
RFC	Request For Comments
SAR	Subject Access Request
SDK	Software Development Kit
SDN	Software Defined Networks
SDR	Software Defined Radio
SHAP	Shapley Additive Explanations
SIP	Session Initiation Protocol
SMA	Software Management Agent
SOC	Security Operation Center
SVM	Support Vector Machine
t-SNE	t-distributed Stochastic Neighbor Embeddings
TCP	Transmission Control Protocol
TEE	Trusted Execution Environment

TID	Telefonica Investigacion Y Desarrollo SA
TUD	Delf University of Technology
UDP	User Datagram Protocol
UE	User Equipment
URL	Uniform Resource Locator
USB	Universal Serial Bus
UT	University of Tartu
VoIP	Voice over IP
VRAN	Virtualized RAN
WBSN	Wireless Body Sensor Network
WebRTC	Web Real-Time Communication
WP	Work Package
XAI	Explainable AI
6LoWPAN	IPv6 over Low-Power Wireless Personal Area Network

1 INTRODUCTION

Artificial Intelligence and especially the field of Machine Learning (ML) has significantly influenced the research and industry in the last years. Many ground-breaking applications have demonstrated the great potential and opportunities of this technology. So far, the main objective in the design and development of AI applications and AI-based systems has been to achieve the highest possible accuracy in the classification and prediction capabilities of the underlying models. However, when AI applications and AI-based systems start to affect and interact with the real world and people (e.g. in the context of cyber-physical systems), other metrics and characteristics have to be considered. These include, for example, the fairness, bias, security, and robustness of the AI models and applications.

Besides the above aspects, the trust of the affected users in the AI must be taken into consideration. Since AI applications are often highly complex non-linear systems, that are perceived as opaque black boxes, it is difficult for users to understand the systems' behaviour and comprehend decisions. However, to gain trust and acceptance - from the users' perspective - in such systems, the systems' explainability and transparency play a crucial role. In the context of the above-mentioned challenges, the European Commission (EC) has defined the goal to enable, accelerate, and regulate the implementation of so-called Trustworthy AI [2]. For this purpose, a high-level expert group on Artificial Intelligence was established that has identified seven guiding requirements that Trustworthy AI systems must fulfil [128]. Precisely, it is recommended that the development, deployment, and use of AI systems meet the following requirements: "*(1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being and (7) accountability*" [2].

The SPATIAL Project addresses these challenges by designing and developing resilient accountable metrics, privacy-preserving methods, verification tools and system framework that will serve as critical building blocks to achieve trustworthy AI in security solutions. The security solutions are investigated in the context of four use cases of industrial and social relevance, namely: (1) the utilization of privacy preserving AI in the cloud-fog-edge continuum, (2) improving the explainability, resilience and performance of cybersecurity in 4G/5G/6G and IoT networks, (3) the utilization of accountable AI in next generation emergency communication and (4) resilient cybersecurity analysis based on machine learning models. These use cases serve, besides an extensive literature review, as the basis for identifying the requirements for the emerging tools and frameworks for privacy preserving and accountable AI/ML in various highly distributed system architectures.

1.1 SCOPE AND OBJECTIVES OF THE DELIVERABLE

This document is part of the first phase of the SPATIAL project, which focuses on identifying requirements and design guidelines[1] for modern system architectures based on accountable AI, as well as proposing resilient accountability metrics and their embedding into the existing AI algorithms. The deliverable document at hand is to be regarded as the basis for achieving these goals. Precisely, this document aims to establish a comprehensive catalogue of aspects and general design principles to be considered when integrating and utilizing AI algorithms and frameworks for addressing security risks and threats to system (and network) architectures. Specifically, in the collection of such aspects and general design principles, we will focus on systems and applications related to the four main technical contexts addressed by SPATIAL: Mobile Edge Systems (e.g. 5G Services), Cybersecurity Applications and Analytics, IoT, and eHealth. The aspects and design guidelines identified in this document are intended to support the design of secure, robust, and explainable AI-based systems. In this context, it is important to notice that although the aim of this document is not to identify concrete system and design requirements for the SPATIAL use cases or the SPATIAL explanatory AI platform, some of the high-level recommendations and guidelines presented here are highly relevant for the later and should therefore be directly taken into account in in the first iteration of the prototype implementation. Furthermore, the recommendations together with experiences gathered from the initial prototyping activities will result in a comprehensive set of final system and design requirements, which will be presented in deliverable D1.3.

To achieve the above goals, we will perform a requirements analysis in the course of this deliverable document. This requirements analysis aims to identify the mentioned relevant aspects and design principles and record them in the form of precise requirements. The foundation of the performed requirements analysis is the strong domain expertise of involved partners, the four SPATIAL use cases, and an exhaustive literature review. As a result of the conducted requirements analysis, the following contributions can be expected from this deliverable document:

- definitions of relevant and necessary terms, which serve as a point of reference for the entire SPATIAL project
- short review of security risks and threats to system and network architectures
- discussion of challenges when integrating AI into modern system architectures.
- identification of stakeholders that are most relevant for the activities and objectives of the SPATIAL project.

[1] This document captures the needs and recommendations for designing AI-based systems which are identified in the first iteration of the SPATIAL project. In this sense, we formulate the general framework of guidelines (written in the form of requirements), which are further refined and specified in later deliverables in the following iterations of the SPATIAL project.

- identification of general aspects and design principles, to be considered when integrating and utilizing AI algorithms and frameworks for addressing security risks and threats to system architectures, in the form of precise requirements
- brief insight into the field of Explainable AI (XAI) and relevant XAI methods that can enhance the explainability of AI-based systems

1.2 LIFECYCLE & RELEVANCE OF THE REQUIREMENTS

This document represents a preliminary catalogue of aspects and general design principles to be considered when integrating and utilizing AI algorithms. These aspects and general design principles were formulated as requirements to be followed and refined during the first prototyping iteration, such that resulting precise system requirements will be reported and summarized in D1.3 "*Final Requirements Analysis for AI towards Addressing Security Risks and Threats to System and Network Architectures*". Along with the other WP1 results, this initial compilation of requirements constitutes the foundation for other technical activities in the SPATIAL project by providing valuable recommendations for designing explainable and accountable AI-based systems. However, to allow to improve the initial design guidelines reflected in the requirements as well as to agilely adapt the technical developments in the project, SPATIAL aims to establish **fully integrative and agile knowledge transfer processes** between the work packages. Precisely, during the course of the project, the requirements will be refined in an agile manner by incorporating the feedback and insights of the other technical work packages into the WP1 activities (see Figure 1). Subsequently, an updated and final set of requirements and design guidelines will be provided to the interested audience in August 2024 (M24) in deliverable D1.3 *"Final Requirements Analysis for AI towards Addressing Security Risks and Threats to System and Network Architectures"*.

The connection and relevance of the requirements to other SPATIAL activities

Figure 1 illustrates the above-mentioned connection and knowledge transfer between the WP1 activities and the other technical work packages. As shown in Figure 1, the design guidelines and requirements provided in WP1 serve as the foundation for the technical activities in WP2, WP3, and WP5. For example, some of the requirements and design guidelines specified in this document are also relevant for the four SPATIAL use cases as these are representative pilots for AI-based systems. Requirements that should be considered in the implementation of at least one of the four SPATIAL use cases have therefore been assigned the relevance *"Relevant for SPATIAL use cases"* (see Appendix A). In this context, we want to recall that each of the four SPATIAL use cases reflects only individual aspects of the SPATIAL key pillars (i.e. privacy, accountability, resilience). Hence, none of the use cases will meet all requirements of this type.

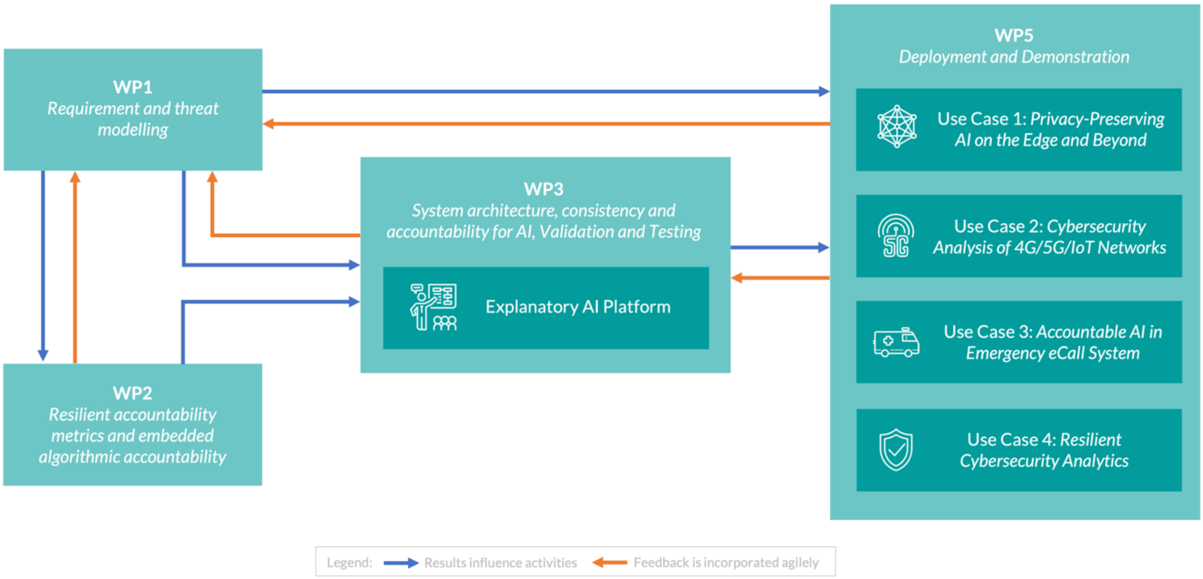
WP1
Requirement and threat modelling
WP2
Resilient accountability metrics and embedded algorithmic accountability
WP3
System architecture, consistency and accountability for AI, Validation and Testing
Explanatory AI Platform
WP5
Deployment and Demonstration
Use Case 1: Privacy-Preserving AI on the Edge and Beyond
Use Case 2: Cybersecurity Analysis of 4G/5G/IoT Networks
Use Case 3: Accountable AI in Emergency eCall System
Use Case 4: Resilient Cybersecurity Analytics
Legend:
Results influence activities
Feedback is incorporated agilely

FIGURE 1: CONNECTION AND KNOWLEDGE TRANSFER BETWEEN WP1 AND THE OTHER TECHNICAL WORK PACKAGES.

Furthermore, the catalogue at hand also includes concrete aspects that are highly relevant for shaping the design and specification of the *Explanatory AI Platform* developed in WP3 (see Section 4). Therefore, we have assigned the relevance "*Relevant for Platform Components*" to these kinds of requirements in Appendix A. These requirements represent high-level recommendations or accountability needs that the platform should address to be beneficial for the four SPATIAL use cases or general AI-based systems. Based on these kinds of requirements, it should be possible to identify required platform functionalities and shape corresponding services. Thereby, it is important to note that a single requirement can have multiple relevance scores assigned to it, meaning it can be relevant for the use cases and the platform at the same time.

Lastly, we assigned the label *"Relevant in General"* to requirements and design guidelines that should be generally considered in the design, specification, and development of AI-based systems. Although they represent relevant recommendations, these requirements do not need to be considered in the context of SPATIAL, as they are either out of scope, not relevant for the use cases in their current TLR level or refer to operational aspects in production environments.

The requirements gathered in this document present just an initial set of design guidelines and recommendations to consider when integrating and utilizing AI algorithms. As the project progresses, we aim to incorporate gained experience and insights from activities of other technical work packages into WP1 in an agile manner[2]. As shown by the orange arrows in Figure

[2] Relevant activities encompass, among others, (a) the definition of concrete resilience and explainability measures/metrics (WP2), (b) research and insights about the accountability and explainability of common

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

1, we aim to use feedback received from these activities and the development of the SPATIAL use cases (WP5) and explanatory AI platform (WP3) to update the listed requirements iteratively and agilely. As stated above, the final set of requirements will be provided in deliverable D1.3. We expect this deliverable to contain requirements and design guidelines that are more specific and tailored to the SPATIAL platform and the use cases but still relevant to a general audience interested in the design of accountable AI-based systems.

1.3 STRUCTURE OF THE DELIVERABLE

The remainder of this deliverable document is structured as follows: In Section 2, we will discuss important and relevant background information which helps to contextualize the deliverable. Specifically, we will define required terms, review security threats for modern system architectures, and discuss challenges in integrating AI into system architectures. Based on this, Section 3 represents the central section of this work. This section will describe the conducted requirements analysis and present the obtained results. Precisely, we will present a stakeholder identification and discuss identified requirements that reflect aspects and general design principles to be considered when integrating and utilizing AI algorithms and frameworks for addressing security risks and threats for system (and network) architectures. In Section 4, we will provide a short outlook on how some of the identified requirements can be used to derive design goals for and required functionalities of the SPATIAL Explanatory AI platform to be designed and developed in context of WP3. To conclude the work, we will summarize the main findings and next steps in Section 5. Finally, we give the interested reader an optional brief insight into the topic of Explainable AI in Appendix B.

AI algorithms and design principles (WP2, WP3, and WP5), (c) assessment of AI-based systems (reliability, limitations, etc) (WP2, WP3, and WP5), (d) metrics for determining what attackers can accomplish and with what resources and capabilities (WP1 and WP2), as well as (e) the technical definition of the SPATIAL platform including the design and specification of a particular distributed AI architecture with its components, properties, processes and flows (WP3).

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

2 BACKGROUND

Before we present the conducted requirements analysis in the next section, we will introduce important and relevant theoretical background information in the following section. This information is fundamental to understand the work at hand and its achieved results and recommendations. First, we will start by defining some relevant terms. Afterward, we will briefly review security risks and threats to system and network architectures. Thereby, to not exceed the scope of this deliverable, we will limit ourselves to 5G/6G networks and IoT and Edge Computing. Subsequently, we will discuss challenges to be addressed when integrating AI into system and network architectures.

2.1 DEFINITION OF RELEVANT TERMS

The subsequent section aims to define terms that are of significant relevance for the present deliverable document and the SPATIAL project as a whole. Therefore, the definitions provided in the following section apply to the entire SPATIAL project and are to be understood as a general point of reference.

2.1.1 ACCOUNTABILITY

Accountability is most widely accepted as 'the obligation to explain and justify conduct' while also raising the warning that 'accountability is elusive' [3]. It roots from Liberation, where it was introduced as a term to limit the use of power, and was later applied for governance. It is especially of importance when the entity in power does not behave as expected, and there is a need to understand the reason behind the action and identify the responsible person or organisation. As a result, it also inherently motivates actors to behave in a better way [5].

Now-a-days a lot of our decision-making tasks are taken by AI, with its implementations in a number of critical applications, from the automotive to the healthcare domain. Thus, its roots are growing deeper into the societal aspects of our lives. Thus, from a socio-technical point of view, AI systems have necessitated accountability as a property for technical systems. As a result, there are changes being made to the governance of AI applications resulting in the legislative changes in General Data protection Regulation (GDPR) and the integration with open government initiatives [5].

The authors in [5] attempt to define algorithmic accountability for AI. For this, they use one of the pioneering definitions [1] which defines accountability as: *"a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences."*. For algorithmic accountability, the actor can be an individual, a group, or an organization responsible for the AI decision making that caused the action. The forum could be legislative committee (e.g., auditing

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

committee) which can ask for clarifications and explanations for the action and can pass their judgement for the actor.

To achieve accountability, the EU Ethics guidelines for Trustworthy AI [2], enlists four main requirements, namely:

- **Auditability:** AI systems should have mechanisms to facilitate auditability, such as ensuring traceability and logging processes. The systems' algorithms, data and design process should be assessable. Systems that affect fundamental rights should be independently auditable.
- **Minimizing and reporting negative impact:** The system should identify, asses, document, and minimise the risks associated to the system, and report misshapen aspects. This could be facilitated by performing risk or impact assessments that consider the different affected stakeholders.
- **Documenting trade-offs:** The above requirements necessitate making trade-offs. AI systems should be evaluated using ethical principles during audits, wherein the necessary trade-offs are explicitly identified, acknowledged, and implemented in AI systems accordingly by the decision maker. If the trade-off is ethically unacceptable, the AI system should not proceed.
- **Ability to redress:** When an unjust adverse impact has occurred, the system should be able to perform adequate redress.

A promising means to achieve accountability is to make the systems transparent by Explainable AI (see Appendix B). Transparency (see Section 2.1.4) can help to achieve accountability but in itself is not sufficient as it is not actively involved in explaining the cause of the output [5]. According to the social scientists, XAI is only a means to explain the causes of an action caused by AI. For achieving accountability, finding an accountable person or organization is also necessary, which requires the study of potentially complex relations of the involved actors [4] [5].

2.1.2 EXPLAINABILITY AND INTERPRETABILITY

In recent years, complex AI-based models have achieved a lot of success due to their impressive capabilities of prediction accuracies. However, these models are often complex nonlinear structures that are highly non-transparent. Therefore, it is non-intuitive to understand how these models reach these decisions. Thus, they are generally regarded as black boxes [6]. Though there is no generally accepted definition for XAI [7], **explainability,** in the context of AI, can be considered to be the capability of understanding the rationale behind the decision-making process of AI [6]. From a human-centric perspective, explainability of an AI model is regarded as the ability for a human to understand the functioning of a decision-making process for a given model with the help of its feature space, training records, targets, and the machine learning algorithm itself. Consequently, the model explanation should be sufficient to inculcate trust in the user about the model including the reasons for the model's behaviour.

As already indicated in the previous subsection, having more explainability would inherently increase the accountability for the decisions made by the AI models since it would be possible to understand the responsible components or features of the model that contributed to the decision-making process. Thus, the obligation appearing through explanations to accept the responsibility would make these models accountable for their decisions.

In contrast to explainability, **interpretability** is referring to a comparatively narrower notion of understanding the cause and effect observed in a black-box model. In other words, interpretability is a mean for providing explainability [8]. Depending on the transparency level of a model, users can decide to utilize tools based on feature/prototype analysis to gain insights of the model's decision-making process.

Although interpretation constitutes a substantial part of an explanation, it's deemed to be insufficient on its own to ensure accountability of a model. In addition, a model must be able to endure scrutiny, offer appropriate explanations to the relevant stakeholders, and validate the decisions it takes, in order to acquire the confidence and accountability of its users and stakeholders.

2.1.3 RESILIENCE

Resilience, which comes from the Latin word "resilio", literally means "to leap back" and denotes a system attribute characterized by the ability to recover from challenges or disruptive events. The resilience concept began to influence other fields such as anthropology, sociology, or psychology, and in the past decades into cybersecurity. In particular, we synthesise various definitions of "resilience" given by the National Institute of Standards and Technology (NIST) [9] as follows: **Resilience is the ability of an information system to reduce the magnitude, impact, and/or duration of disruptive events or unknown changes in the operating environment (including deliberate attacks, accidents and naturally occurring threats or incidents) by *a)* anticipating and preparing for such events (e.g., through risk management, contingency and continuity planning); *b)* being able to withstand and adapt to attacks, adverse conditions or other stress and potential disruptions, and continuing to operate while maintaining essential and required operational capabilities; and *c)* recovering full operational capabilities after such a disruption in a time frame consistent with mission needs.** Cyber-physical systems (CPS), including IoT, Edge, and 4G/5G systems, are systems that tend to evolve on a larger scale and have considerable dynamics, complexity, and heterogeneity of complex components. For these systems, the term "resilience" is about the ability to resist, absorb, recover, or successfully provide the essential services and keep an acceptable performance even under failures or crises (e.g., cyber-attacks or security breaches).

AI models have been employed in different critical components of CPS systems. The resilience of AI-based cyber-physical systems against attacks and other environmental influences needs to be ensured like for other assets. The concept of "**resilient AI**" means that the AI-based systems continue to offer the services they should in the presence of adversarial attacks (e.g., evasion/poisoning attacks) to guarantee safety and security of the systems where they are deployed.

2.1.4 TRANSPARENCY

In common daily life systems, transparency deals with hiding differences about how data is processed and exchanged between components, such that it can be presented to the end user in a coherent form on the screen [14]. Applications accessible through the Internet provide a high degree of transparency to end-users. Indeed, the functionality of applications is typically distributed across multiple computers that can be also deployed in different locations. This distributed complex functionality is presented and perceived to the end-user in an invisible manner, e.g. via a web browser.

In contrast to hiding complexity from the users, **transparency** in the case of systems that implement AI models for decision making focuses on dissecting the complex logic involved in the decision process of a model, such that it is possible to determine the causes that lead to that decision [12]. The right of obtaining such explanation from the users is supported by the fact that models work in a black-box manner and can be easily biased by programming errors from developers and quality of the data used in the training process [12]. Thus, transparency fosters a way to ensure fairness in an automated complex AI system for individuals [11].

Several types of explanations ranging from numerical values to phrases can be used to explain transparency of a model [16]. Transparency may also provide insights into characteristics of the system that can be hard to measure - for instance, fairness and causality. Naturally, how transparent a system is depends on the method to explain how the system works even when it behaves unexpectedly. This can be difficult to achieve in practice, even if the source code of the application is available for analysis and audit. For example, complex structures as neural networks are not really possible to explain in detail as it is difficult to prove that they work correctly on new unseen observations [15].

2.1.5 TRUSTWORTHINESS

The term *trustworthiness* is defined by the National Institute of Standards and Technology (NIST) as *"the attribute of a person or enterprise that provides confidence to others of the qualifications, capabilities, and reliability of that entity to perform specific tasks and fulfill assigned responsibilities"* [19]. Similar to this definition, Mohensi et al. describe trustworthiness as *"positive user attitude toward the system that emerges from knowledge, experience, and emotion"* [17]. Achieving trustworthiness in the context of XAI is not a trivial task and relies on three main components identified by the European Commission as *(i)* lawful, *(ii)* ethical, and *(iii)* robust [20]. This means that the AI system has to comply with the established laws, regulations, and ethical norms. At the same time, the AI system has to operate in a reliable and safe manner both from technical but also social perspective [20].

Within the scope of the above definitions, an AI model can be considered *trustworthy* when the end-user can trust that the model reliably generates unbiased and fair predictions. The model trustworthiness can be judged with regards to model fairness, reliability, and safety [17]. Fairness refers to the ability of the model to perform a fair feature learning without any bias towards a specific label or feature class [17]. On the other hand, *safety* and *reliability* refer to the model ability to learn the input features and the corresponding labels in a robust manner, which covers the ability of the model to operate only within the constraints of its intended functionality. This means that a trustworthy model is expected to preserve the confidentiality, integrity, and availability of the input data it operates on [19]. Additionally, the model is expected to generate accurate predictions without failures even in the presence of external risk factors (e.g., environmental disruptions, human errors, purposeful attacks, etc.) [19].

In addition to a trustworthy model, a second level of trustworthiness in XAI can be established when users put trust into individual predictions and take actions based on that trust [18]. This can be defined as *"prediction trustworthiness"* and is not limited to predictions generated by trusted (and trustworthy) models. Users can trust individual prediction given enough reasoning behind the decision making of the model when generating the prediction. However, the user is not obliged to trust the model as a whole despite trusting an individual prediction coming from it. The relationship between prediction and model trustworthiness is independent.

2.1.6 RISKS

The term *risk* describes the possibility/potential for unexpected, possibly harmful, or dangerous event taking place. The term covers not only the probability for experiencing a loss or a negative event, but also the extent to which certain assets or system infrastructure will be damaged [21]. The NIST defines the term as "*A measure of the extent to which an entity is threatened by a potential circumstance or event, and typically a function of: (i) the adverse impacts that would arise if the circumstance or event occurs; and (ii) the likelihood of occurrence.*" [23]. Cherdantseva et al. propose a similar way of calculating risk - as a product of the loss from the attack and the probability for the attack [22]:

Risk = (probability of attack) x (impact/loss of the attack)

This formula illustrates that even very unlikely to occur attacks can lead to a substantial risk when the potential loss caused by the attack is also very large [22]. Security risks can be classified as both information- and system-related security risk and typically involve confidentiality, integrity or availability loss, privacy or financial loss, present legal implications, physical- and health-related losses and even image and reputation loss [23].

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

2.1.7 THREATS

A *threat* can be defined as any circumstance or event with the potential to adversely impact the *assets* of an organization through the information system. Security threats cause damages to the asset's *security attributes* such as confidentiality, integrity, and/or availability. They exploit *vulnerabilities* that can be qualified as physical, technical, or administrative weaknesses of the information system. Each threat can be characterized by its *origin*, which can either be human or natural (e.g., earthquake, thunderstorm, fire, etc.). It is also characterized by its *source*, which can either be accidental (e.g., hardware failure, power outage) or deliberate (e.g., alteration of software, theft). A threat can give rise to other threats and thus, it can also be a consequence of another threat, generating the concept of *threat chains*. Figure 2 depicts a taxonomy of threats in interaction with the attributes they interact with [24].

When talking about threats in a security context, we mostly focus on human and deliberate security threats: the ones rising from attackers wanting to compromise the system. The identification and quantification of threats is paramount to manage information security in organization. Together with vulnerability knowledge, threat knowledge enables to quantify the real-world security risk of an information system as well as the prevalence of threat conditions if a vulnerability is exploited and materializes into an actual attack. Security threats can be identified through exercises called threat analysis or threat modelling. The prevalence of a threat can be quantified in different manners. For instance, statistics of observed security incidents exploiting specific vulnerabilities can be used to empirically quantify the prevalence of a threat. Alternatively, proxy indicators like the level of interest of an attacker for compromising the system and the ease to exploit a vulnerability can be used to quantify a threat. The knowledge about security threats is typically managed on a global scale through threat intelligence that tracks how vulnerabilities are being exploited by threat actors around the world.

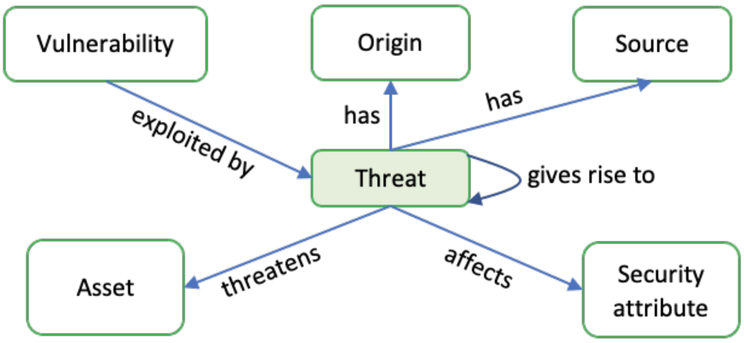

FIGURE 2: TAXONOMY OF THREATS IN INTERACTION WITH THE ATTRIBUTES THEY INTERACT WITH [24]

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

2.1.8 PRIVACY

In SPATIAL, we ground the definition of *privacy* on the discussion in GDPR. The GDPR importantly considers two elements in the definition of data privacy: the entity requiring data access and the purpose of the access. In addition, the GDPR summarizes the key concept of data privacy as "*empowering the users to make their own decisions about who can process their data and for what purpose.*" [25]

An important concept that is coupled with data privacy is *personal data*. Again, GDPR provides a broad definition of it, that is, "*any information which are related to an identified or identifiable natural person.*" By having a very open definition through the phrase, "*any information*", GDPR not only recognizes the information that can identify a natural person as personal data (such as address, identity number, etc.) but also any data generated by the person (such as for example phone call logs, website access logs and purchase history), and also any meta-data generated from such data (such as patterns extracted from the logs, labels assigned to the person by a machine learning algorithm etc.) [26].

2.2 REVIEW OF SECURITY RISKS AND THREATS TO SYSTEM AND NETWORK ARCHITECTURES

The following subsection will review security risks and threats to system and network architectures. This is expected to provide essential insights that may be of particular relevance to the recommendations and general design principles derived in this document. However, to not exceed the scope of this deliverable, we will limit the review to 6G networks and IoT and edge computing systems.

2.2.1 AI ATTACKS ON 6G RAN

Radio Access Networks (RAN) are telecommunications components that connect mobile devices/user equipment (UE) to a public or private core network over an existing network backbone. It is possible to provide ultra-reliable (deterministic) wireless performance with Long-Term Evolution (LTE) and 5G RANs[3], which are expected to further improve in 6G. As 2G, 3G, and 4G radio access technologies have progressed, new RANs such as GERAN, UTRAN, and E-UTRAN have been developed to handle the increased data demands of these newer mobile technologies. Recent enhancements, such as Centralized/Cloud RAN (CRAN), Virtualized RAN (VRAN), and Open RAN (ORAN), are projected to be linked beyond 5G/6G and include other modern technologies such as Software Defined Networks (SDN), and Network Function Virtualization (NFV).

[3] Base-band units, radio units, remote radio units, antennas, and software interfaces make up a RAN.

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

Security threats in 6G RAN

Previous studies including the ORAN alliance, have reported attacks such as denial of service (DoS), eavesdropping, man-in-the-middle (MITM) attacks, Medium Access Control (MAC) spoofing, identity-theft attack, jamming attacks, and Transmission Control Protocol (TCP)/User Datagram Protocol (UDP) flooding among many other. However, there are some threats which are inherited from the predecessor of CRAN, Cognitive Radio Networks (CRN). For example, Primary User Emulation Attack, Spectrum Sensing Data Falsification attacks, Common Control Channel attacks, Beacon Falsification attacks, Cross-layer attacks targeted at several layers, and Software Defined Radio (SDR) attacks are to name a few.

Attacks on 6G RAN AI Solutions

Machine Learning based Intrusion Detection System (ML-IDS) is the most promising anomaly-based Intrusion Detection System (IDS) proposed in studies because of their capability to gradually learn new information while performing a given task. Authors of [27] have used multi-layer perceptrons (MLPs) and Support Vector Machines (SVMs) enabled with the *kernel trick* to classify and detect multi-stage jamming attacks in CRAN Baseband Unit (BBU) pool. ORAN is envisaged to be the future of RAN technologies in Beyond 5G (B5G). When it comes to ORAN, self-organization and intelligence-based technologies will be extensively used in the deployment process [28].

2.2.2 IOT AND EDGE COMPUTING

Edge computing is an attractive approach to addressing latency and bandwidth demands of emerging applications by moving computing services closer to the source of the data or end devices at the edge of the network, such as IoT devices. Therefore, the main security issue of edge computing is having a large attack surface and exposure to threats, including data leakage, distributed denial of service (DDoS) attack, and intrusions into local networks or cloud resources.

In computing, a DDoS attack is a common and very powerful attack, in which an attacker aims to violate the availability of the target server or infrastructure by making it unavailable to its users with a massive amount of traffic, such as in the Domain Name Server reflection attack [29]. From January to June in 2021, as Kaspersky reported [30], some 1.51 billion breaches of IoT devices took place, an increase from 639 million in 2020. Most attacks compromised the remote communication between IoT devices and the server using Telnet protocol attacks. Concretely, "*more than 58% of IoT cyberattacks leveraged the vulnerable protocol Telnet with the major intent of cryptocurrency mining, DDoS shutdowns, or pilfering confidential data*", according to Kaspersky [30]. Furthermore, in the past, most IoT attacks were driven by two prominent IoT botnets: Mozi and Mirai. While the older botnet Mirai appeared in 2016 was unleashed massive DDoS attacks on major websites using millions of compromised devices, Mozi, which is a Mirai botnet variant, has been extremely active in 2019. "*The Mozi botnet controlled approximately 438,000 hosts, which is determined by the count of unique Mozi Uniform Resource Locators (URLs)*", according to the tracking records of [31]. Recently, the newest Mirai-type variant in 2022 has been discovered targeting both known and zeroday vulnerabilities in D-Link, SonicWall and Netgear devices, as well as in unknown IoT devices.

In December 2021, a remote command execution vulnerability CVE-2021-44228 was disclosed in Apache Log4j, which is a highly popular, widely deployed logging tool used in many big companies such as Google, Apple, Steam, etc. This vulnerability is one of the most critical vulnerabilities in the history of modern edge computing as *"more than 2.5 billion devices running Java, coupled with the fact this vulnerability is extremely easy to exploit, consequently the impact is very far reaching, as the Heartbleed vulnerability and Shellshock combined*", according to [32]. Technically, an attacker who can control log messages or log message parameters can execute arbitrary code loaded from Lightweight Directory Access Protocol (LDAP) servers when message lookup substitution is enabled. This vulnerability could therefore allow even unsophisticated threat actors to take remote control over millions of IoT and other edge computing devices as well as applications dedicated to the enterprise.

2.3 CHALLENGES WHEN INTEGRATING AI TO SYSTEM ARCHITECTURES

Due to achievements in the availability of data and computational resources, AI applications play an increasingly important role in many domains [33]. As a result, AI is often integrated into traditional system and network architectures, and AI-based systems are becoming more pervasive [34]. However, since traditional systems often follow different software and system engineering processes and design patterns, integrating AI components poses new challenges [35]. Therefore, building and maintaining AI-based system architectures is not a trivial task, and many diverse aspects need to be considered [34]. Hence, in the following, we will present the most recent challenges when using or integrating AI into system architectures. In this context, we follow the definition for AI-based systems of Martínez-Fernández et al. [34], who define an AI-based system as "*A system consisting of various software and potentially other components, out of which at least one is an AI component.*".

Supply of high-quality Data

Data and its management are among the most important aspects to consider for AI-based systems and the often-utilized machine learning algorithms. As the fuel of AI algorithms, the supply of a large amount of high-quality data is essential for the successful development of data-centric AI applications. Since data shapes AI models, their quality depends primarily on the quality of the supplied data, and the behaviour of AI-based systems becomes data-dependent [33] [36]. However, the collection of a large amount of high-quality data is not a trivial task and

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

poses many challenges. First, AI applications require a large amount of data. Therefore, the biggest challenge is to collect a considerable enough amount of data that is sufficiently representative for the problem to solve. Second, it is also necessary to ensure that the gathered data is fair, unbiased, and balanced. In addition, it should be ensured that the data contains sufficient rare cases. This is the only way to ensure that the AI system meets ethical and legal compliance requirements, makes fair and unbiased decisions, and does not reinforce existing discriminations [34].

Another typical challenge when using AI is the costly labelling of data [34]. Many ML models require labelled data during the training phase. Creating a labelled data set from real-world data is often very expensive since the data typically needs to be analysed and labelled by human domain experts. These cost factors must be considered when integrating AI into systems.

In addition to collecting data, the processing and management of data are also challenging since they often involve the definition of complex and dynamic processes. On the one hand, especially in the IoT area, many data originate from heterogeneous and diverse data sources, leading to various data dependencies [35] [36]. Furthermore, merging and homogenizing such diverse data poses additional challenges regarding data wrangling, handling big data streams, and defining appropriate data structures. In addition, the management and analysis processes such as data exploration, data cleansing, and data analyses also pose challenges. It must be ensured that these processes can be integrated into a dynamically adjustable pipeline within the system, in order to be able to adapt to changes quickly.

The last data-related challenge to mention is that the data supplied in the deployment phase must be continuously monitored and reviewed [33] [36]. Similar to the real world, the statistical distributions represented in the training data change dynamically [37]. This can result in a slow degradation of the performance of the data-centric AI-based system due to the dynamic changes in the supplied production data. Therefore, the supplied data must be regularly analysed, extended, and in case of a distribution drift, the AI model must be adapted. To summarize, the collection of high-quality, fair, and unbiased data and their continuous monitoring require the implementation of additional data processing and monitoring mechanisms, which makes the integration of AI in the existing system architecture more challenging.

Trust Deficit and the need for Explanations

In the last decade, deep learning applications like AlphaGo [38] have demonstrated superior performance to humans in the game of Go. However, the example of AlphaGo proves that an increase in performance comes with an increase in complexity which in turn results in less explainability. Even the best domain experts are often no longer able to understand the decisions and highly non-linear complexity of the high-performance AI models. Therefore, most AI models are perceived as a black box, opaque, non-intuitive, and difficult for people to understand. However, understanding the system's functioning is a crucial driver of trust [36]. Furthermore, when delegating control to systems, trust is also closely related to the correct, safe, and unmalicious behaviour of these systems [37]. Therefore, due to the opaque and uncertain decision-making process, AI/ML practitioners and the end-users may not fully trust AI-based systems, and a trust deficit can arise.

In addition, the uncertainty in the decision-making process and the black-box computations pose enormous challenges, especially when integrating AI into system architectures. In practice, AI models are not deployed as a standalone piece of software, and instead, they are surrounded by multiple components with which they form a complex computational system. Since the ML/AI model generates predictions in a black-box manner, this can lead to unexpected system behaviour in the production environment, where potentially untested circumstances and data inputs can be forwarded to the AI system. In such cases, the AI/ML practitioners and the end-users must trust that the system does not perform any malicious activity, despite being a black-box. However, especially when using AI in safety-critical applications, blindly trusting an AI/ML model is not acceptable. The rationale behind the decision-making needs to be provided, and users need to know, in which circumstances the system is trustful and when not [36]. This clearly elucidates the need for interpretability, explainability, and even accountability of AI models when integrating them into system architectures.

Security and Privacy Issues

Similar to any other software system, the security of AI-based systems is of utmost importance. In this context, it is crucial to ensure that both traditional and AI components are designed to be secure and prevent the exploitation of known vulnerabilities. Thereby, it must be guaranteed that AI models are at least as secure, robust, and resilient to attacks as the rest of the system, for which traditional security mechanisms already exist. However, testing and reasoning on the security and robustness of AI models is a challenging task [34] [39], and many known attack vectors exist [36] [40] [41] [42]. Especially in the field of machine learning, two broad categories of attacks exist - *poisoning attacks* and *adversarial attacks* [45] [46] [43]. Under the assumption that an attacker has (partly) access to the training data, poisoning attacks try to invade the training phase of ML models by manipulating the supplied training data [45]. The goal of these attacks is to significantly degrade the model performance and move the decision boundary of the targeted model [41]. Prominent examples of poisoning attacks are the label flipping, input manipulation, or data injection attack [43]. As the name indicates, *label flipping attacks* aim to manipulate the target label of specific training instances in order to induce misclassification of distinct samples [40]. In contrast, *input manipulation* and *data injection attacks* aim to tamper the trained model by generating a drift in the training data distribution by either manipulating features of existing data instances or adding new instances with manipulated features [48]. Other than poisoning attacks, adversarial attacks are typically applied at the inference time of ML models [43]. The most prominent example are *evasion attacks*, in which attackers aim to manipulate data instances so that an already trained AI model misclassifies them during inference [41]. Such manipulated data instances are called *adversarial examples* [41]. The goal of evasion attacks, for example, is to use adversarial examples to undermine AI-based security

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

mechanisms like intrusion detection systems and keep active attacks undetected [43]. Furthermore, specific variants of adversarial attacks can be used to gain information about internal model parameters or even the supplied training data. This can raise serious privacy issues and may result in the leakage of sensitive information. For example, attackers can apply *membership interference attacks* to infer whether a specific data point belongs to the training set [41]. Even more harmful from a privacy perspective, *model inversion attacks* enable attackers to extract and reconstruct the training data from the model predictions at inference time [43]. In contrast, adversaries use *model extraction attacks* (sometimes known as *model stealing attacks*) to extract and recover many parameters and details of attacked black-box models [43]. Precisely, attackers use input-output pairings to approximate a surrogate model that closely approximates the model under attack [41]. Since ML models can memorize training data and need to be considered intellectual properties [43], this also has profound privacy implications. In conclusion, to be secure and robust against these diverse attacks, AI-based system architectures have to support monitoring and security mechanisms that protect the deployed AI model, detect adversarial attack attempts, and provide measures to protect the privacy of sensitive data. In this context, privacy-preserving AI approaches like federating learning or differential privacy should be considered in the very early design stages of AI-based systems [36] [41] [43].

An additional challenge that raises serious privacy concerns is related to the resource-intensive training process of ML models [37]. Since the machine learning training process is highly computational-demanding, the training is often outsourced to third-party cloud service providers to avoid the acquisition of specialised hardware. In this situation, it must be guaranteed that either no sensitive data is supplied to the third-party service provider or strict processing agreements that comply with EU data protection regulations are signed. Further privacy-concerning aspects to be considered are the often undeclared consumers of AI model predictions [35]. In complex AI-based systems, the prediction of models is often provided as input to other components or even written to logs. In this case, possible leakage of sensitive data must be avoided at all costs. Being aware of data protection regulations and consumers of AI predictions is a challenge when integrating AI into system architectures.

Model and System Evaluation, Maintainability, and Adaptability

Evaluating AI models and AI-based systems is a challenging task. Many benchmark metrics either exist for accessing the performance of AI models or traditional systems. However, there is a lack of metrics and performance benchmarks for conjunct AI-based systems [37]. Even more challenging, there are many conflicting quality attributes, for which traditional and AI-based systems are designed [34]. Typical quality attributes for traditional systems are response time, scalability, robustness, safety, and security. For AI models, it is hard to reason about their robustness, safety, or security attributes [34]. Instead, the mere model performance is typically used as a quality attribute for AI applications. Depending on the specific problem class (e.g., classification vs. regression), performance metrics such as classification **accuracy**, **false/true**

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

positive rate, **F1 score**, or the **mean squared error** are used. In this context, it needs to be mentioned that in order to compare two or more AI models properly, they have to be trained and validated on the same train, validation, and test datasets. Additionally, the same evaluation metric must be used for all models. In summary, a trade-off regarding these partly disparate quality optimization goals between AI models and traditional systems has to be found when designing AI-based systems.

Further challenges closely related to the model performance are the maintainability and adaptability of AI-based systems. As mentioned above, the statistical distribution of the data supplied in the deployment stage can deviate from the utilized training data. Besides, even novel and more representative data features can emerge. As a consequence, the performance of the corresponding AI model can drop significantly. Therefore, the performance of the AI model needs to be continuously monitored in the production environment [33]. If any data distribution deviations or model performance decreases are detected, it is necessary to rapidly adjust and retrain the model in the run-time context [37]. In this context, the AI-based system has to allow models to evolve iteratively, and the system needs to be able to execute model training, selection, and deployment quickly. The designed system has to enable rapid and isolated model adaptions without requiring the adjustment of other system components [34]. In conclusion, AI-based systems require a robust and evolutionary infrastructure [34] that supports updating the model in a rolling manner. Thereby, deciding when to perform model retraining is a challenge itself.

Testing and the Difficulty of Certification and Standardization

Testing and quality assurance of AI-based systems is a multi-dimensional and demanding challenge. As mentioned above, the decision-making process of AI components is always associated with a certain degree of uncertainty. This is mainly due to the fact that the AI/ML models and the inherently incomplete data cannot represent the full semantic multimodal complexity and complex non-linear relationships of problems [34] [36]. Therefore, due to the lack of absolute certainty, complete coverage of all test scenarios based on inherently limited data is not feasible. Furthermore, most AI/ML models are based on complex probabilistic optimization and learning processes. As a result, these models cannot be strongly specified a priori, rendering their validation and verification difficult [34] [39]. In addition, the opaque black-box computation is an obstacle in testing and debugging AI-based systems. More precisely, it is difficult for testers to identify and fix errors due to the AI models' weak interpretability and explainability [34]. Furthermore, it seems infeasible to formally verify and validate AI-based systems due to the missing testing capabilities and the weak a-priori specification discussed above. In addition, these shortcomings make it very difficult to define and verify general standards for AI-based systems. As a result of both facts, receiving certifications for AI-based systems from official certification bodies is a very challenging undertaking. Especially in the EU, this poses enormous problems for operators of safety-critical AI-based systems in terms of legal approval and accountability.

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

Interoperability, Heterogeneity of ML Models and Codebases, and Pipeline Orchestration

Integrating AI components in traditional complex systems often results in components with heterogeneous codebases written in different programming languages [34]. As a consequence, the dependency management of AI-based systems becomes challenging. Guaranteeing the interoperability of components is not trivial, and strong inter-component entanglement emerges [34] [35]. This results in the *Changing Anything Changes Everything* (CACE) principle [34] [35], which renders the isolated improvement of individual components challenging. In addition, the development process of AI models differs from those of traditional software systems. Typical ML tasks comprise multiple different stages of different types (e.g., data loading, data joining, feature extraction, feature transformation, model training, evaluation, and deployment). This leads to the challenge of orchestrating an interoperable pipeline that implements all these heterogonous stages. In practice, developers often apply several anti-patterns, which result in complex dependency management, glue code, pipeline jungles, dead experimental code path, or configuration debt [34] [35] [37]. Given the heterogeneous nature of each task, it is a substantial challenge to orchestrate the execution of these in a fluent, interoperable, and coherent way. This raises concerns from a system architecture perspective since the architecture must support such pipeline stage integration. Furthermore, the development of AI and ML systems often involves the employment of third-party libraries and frameworks. The integration and optimal utilization of these require detailed knowledge and expertise. Besides, the evolution of these libraries needs to be tracked to avoid incompatible software updates [36]. Using third-party software components also amplifies the software security risks, which must be considered in the very early stages of the software and system design process [34].

Human-in-the-loop settings

Many AI-based systems are designed to support human experts in their decision-making process. In particular, this pattern is often applied in the cybersecurity domain. Such complex and hybrid automated decision systems link AI-based and human-based actions. The AI system is designed as improved assistance for the human operator and provides them feedback in the operational context [36] [44]. The advantage of these hybrid approaches is that humans are more flexible, provide more rapid judgment on security alarms, and can deliver immediate professional judgments. Thus, by combining human expertise and the excellent anomaly detection capabilities of AI systems, the quality and robustness of security systems can be further enhanced [45]. However, such hybrid systems introduce different challenges and requirements. In this human-centric AI setting, the AI system must not just make decisions but also provide complex explanations and auxiliary data. This is important for the human operator to understand the decision-making, judge the current risk situation, and apply appropriate countermeasures. In this scenario, interpretability and explainability again play a crucial role.

Besides the above aspects, the quality, accuracy, and reliability of model predictions are of utmost importance. On the one hand, the detection error rate needs to be minimized. On the other hand, frequent redundant human interaction needs to be avoided so that human analysts are not burdened unnecessarily and lose trust in the AI system. Finding a trade-off between these two goals is a demanding requirement. Finally, it must be mentioned that it is not trivial to evaluate and estimate the performance and reliability of human-centric AI-based systems. In particular, it is not easy to evaluate the human influence and the quality of the human-supporting auxiliary information provided by the AI system. At the moment, there are no widely accepted benchmark metrics for such systems, and further research effort in this direction is required [44].

Hardware Limitations

The training of many of the AI/ML models is based on complex incremental optimization algorithms. As already mentioned, these training processes require a large amount of data and are highly time and resource-consuming [37]. As a consequence, the training of AI models is often employed to specialized hardware or third-party cloud services or distributed on numerous machines (e.g., federated learning) [37]. Therefore, using specialized hardware and highly optimized software is necessary for AI-based system architectures.

In addition, some AI systems rely on massive amount of sensor data generated by numerous IoT devices [46]. In an IoT-based system, some computations have to be performed at the Edge. This poses serious challenges for ML/Deep Learning (DL) workloads due to the limited hardware capabilities of the typical Edge devices. The most common IoT hardware constraints in the context of ML are related to computational performance, accuracy, energy consumption, cost, throughput, and errors in collected data [46] [47]. Finding a balance between these constraints involves a trade-off, e.g., higher throughput will probably lead to higher cost and energy consumption. Therefore, to manage an ML-based system, a certain set of design decisions with respect to the system architecture have to be considered.

3 REQUIREMENTS ANALYSIS

The following section describes the structure, scope, and outcomes of the performed requirements analysis that aims to capture aspects and general design principles to be considered when integrating and utilizing AI algorithms and frameworks. First, we will discuss the general approach of the performed requirements analysis. Afterward, we will examine the scope and limitations of the conducted analysis by presenting the four SPATIAL use cases, which limit the scope of the former. Finally, we present the outcomes of the conducted requirements analysis by discussing the identified stakeholders and captured requirements.

3.1 GENERAL APPROACH AND METHODOLOGY

The aim of this deliverable is to establish a comprehensive catalogue of aspects and general design principles to be considered when integrating and utilizing AI algorithms and frameworks for addressing security risks and threats for system (and network) architectures. In this context, the compiled catalogue is intended to support the design and application of AI algorithms in an explainable manner, which enhances the accountability of AI-based systems. Following the strategy of the SPATIAL project, the aspects and design principles gathered in this deliverable will be restricted to the four domains addressed in the SPATIAL project: Mobile Edge Systems (e.g., 5G Services), Cybersecurity Applications and Analytics, Internet of Things, and eHealth.

3.1.1 COLLECTING REQUIREMENTS AND IDENTIFYING STAKEHOLDERS

In order to achieve the goals mentioned above, a requirements analysis was performed within the scope of this deliverable. This requirements analysis aims to identify the mentioned relevant aspects and design principles and record them in the form of precise requirements. The high-level structure and the resulting outcomes of the performed requirements analysis are illustrated in Figure 3. As can be seen, the collected requirements originate from various inputs. One primary source of the gathered requirements is the strong industrial domain expertise of the involved consortium partners. In addition, the four SPATIAL use cases that reflect the four technical contexts focused on in the SPATIAL project (Mobile Edge Systems, Cybersecurity Applications and Analytics, IoT, and eHealth) play an essential role in the collection of the aspects and general design principles when integrating and utilizing AI algorithms and frameworks (see Section 3.2). Lastly, based on a comprehensive literature review, further requirements and recommendations were identified that are of particular relevance in the context of SPATIAL. The outcomes of the requirements analysis are also depicted in Figure 3. On the one hand, we identified important stakeholders that are most relevant for the activities and objectives of the SPATIAL project (see Section 3.3). On the other hand, we identified and discussed several distinct requirements and classified them into multiple categories (see Section 3.4). Based on these results, we summarize the most important findings and derive the next steps in Section 5.

SPATIAL project is funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 101021808.

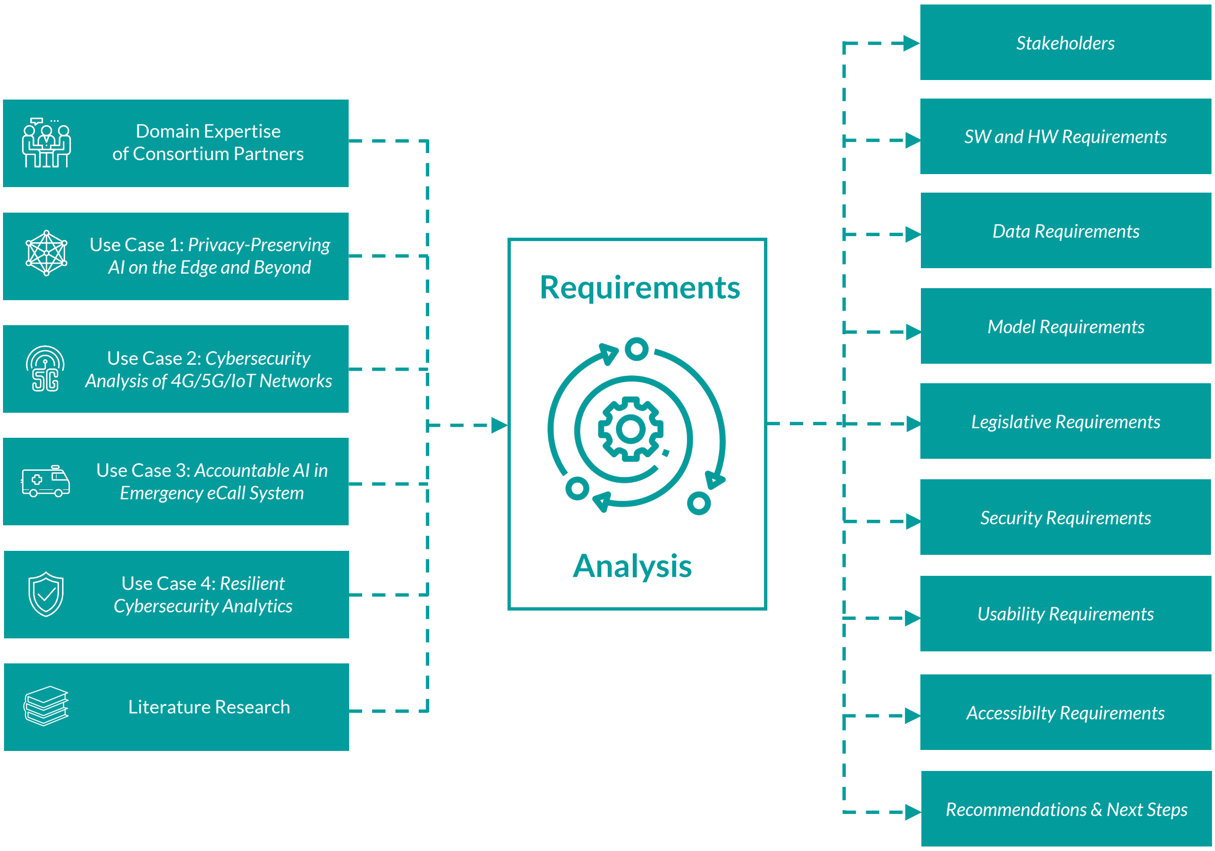
Domain Expertise of Consortium Partners
Use Case 1: *Privacy-Preserving AI on the Edge and Beyond*
Use Case 2: *Cybersecurity Analysis of 4G/5G/IoT Networks*
Use Case 3: *Accountable AI in Emergency eCall System*
Use Case 4: *Resilient Cybersecurity Analytics*
Literature Research
Requirements
Analysis
Stakeholders
SW and HW Requirements
Data Requirements
Model Requirements
Legislative Requirements
Security Requirements
Usability Requirements
Accessibilty Requirements
Recommendations & Next Steps

FIGURE 3: OVERVIEW AND OUTCOMES OF THE PERFORMED REQUIREMENTS ANALYSIS

We also provide a concise classification between functional and non-functional requirements during the conducted analysis. For this purpose, a dedicated column illustrating this mapping can be found in the requirements tables provided in Appendix A. As mentioned above, the requirements gathered in this document not only cover aspects relevant for the technical activities of the SPATIAL project but also represent aspects and design principles that should be considered when aiming to develop secure and trustworthy AI-based systems related to the four main technical contexts addressed by SPATIAL: Mobile Edge Systems (e.g. 5G Services), Cybersecurity Applications and Analytics, IoT, and eHealth.

Hence, besides aiming to extend the knowledge towards realising secure and trustworthy AI-based systems in these four domains, the collected requirements are relevant for the technical activities of the SPATIAL project and provide the general framework for prototyping and deriving specific system requirements in the following D1.3. On the one hand, the requirements are essential for the four SPATIAL use cases as these are representative AI-based systems from the four domains mentioned above. Therefore, the gathered requirements set the direction of the developments and technical activities of the pilots in the context of SPATIAL, as they allow the use cases to shape their work towards achieving a secure and accountable implementation. On the other hand, some requirements also illustrate the needs and technical challenges of AI-based systems when addressing explainability and accountability. Hence, as the SPATIAL

Explanatory AI platform aims to enable explainability and accountability by providing methods that allow stakeholders to understand how an AI system makes predictions, the requirements clearly outline which aspects the platform should address to enable this. This allows the planned technical activities in WP3 to be more targeted towards being beneficial for AI-based systems that aim to address explainability and accountability (see Section 4 for more details).

To avoid confusion and to clarify the relationship between the last two points, we identify for each of the collected requirements (1) whether it is relevant to the technical activities in SPATIAL or just contributes to extending the knowledge towards trustworthiness in AI, and (2) in case the requirement is relevant, by which technical activities it is considered/addressed. For this purpose, a column *"Implemented by"* can be found in the requirements tables provided in Appendix A. This column expresses whether a requirement is taken into account during the implementation of the platform and/or the four use cases. In this context, we want to emphasise once again that this document represents a catalogue of initial requirements. This catalogue will be refined and the requirements will be concretized towards SPATIAL's technical activities in the scope of Deliverable D1.3. This will be achieved by agilely incorporating feedback and insights obtained from the development of the four SPATIAL use cases (WP5) and the Explanatory AI platform (WP3) (see Figure 1).

The process of identifying requirements and stakeholders

The precise process of identifying and collecting requirements was distributed within the consortium and discussed in numerous virtual meetings. As of date 30/04/2022, already ten virtual WP1 meetings have been conducted regarding the organisation and execution of the requirements analysis. At the beginning of this process, relevant requirements groups and stakeholders were collaboratively identified in these meetings. Subsequently, individual requirement groups were assigned to contributing partners. Thereby, it was ensured that this assignment corresponds to the personal domain expertise of the allocated partners. Afterwards, the individual responsible partners identified initial design guidelines in the form of requirements by (1) considering their strong domain expertise, (2) reviewing relevant literature, and (3) contacting and asking the use case leaders, which are representatives of some of the identified stakeholders. This approach guarantees that both the SPATIAL use cases as well as the partners' strong and diverse domain expertise are reflected in the gathered requirements and design guidelines. During the performed analysis, the requirements were collected in a shared space and thus made available to all partners for verification. All WP1 partners involved were asked regularly to validate the collected requirements and incorporate feedback. In this way, the requirements were strengthened across the partners' diverse domain expertise.

As the project progresses, it is planned to refine the initial requirements in collaborative workshops towards Deliverable D1.3. As mentioned above, practical experience and feedback gained from the technical activities of the SPATIAL project, especially from the implementation of the use cases and the Explanatory AI platform, will be incorporated into the refinement of the requirements (see Figure 1).

3.1.2 PRIORITIZATION OF THE GATHERED REQUIREMENTS

All requirements collected in this document comprise a short and precise definition, a unique identifier, and a priority level. For the definition of the priority levels, we follow RFC2119 [68]. This specification defines the following terms to be used for prioritizing requirements[4]:

MUST. The RFC2219 specification describes this term and its synonyms *REQUIRE* and *SHALL* as an *"...absolute requirement...* " [68] that is mandatory to be fulfilled. Therefore, all requirements, aspects, and design principles defined with these priority levels should be strictly realized when integrating and utilizing AI algorithms and frameworks.

MUST NOT. In analogy to the previous term, this phrase indicates that a requirement denoted by this term must be avoided at all costs. SHALL NOT is also used as a synonym in this context.

SHOULD. According to RFC2119, this term defines a requirement as a recommendation that should be implemented if possible. However, the realization of this kind of requirement is not mandatory, and there may be good reasons why a requirement specified with this term cannot or should not be realized. Therefore, the synonym RECOMMENDED is also used.

SHOULD NOT. Similar to the previous term, this word indicates that an aspect or requirement should not be fulfilled. Again, there may be reasons to ignore this and still realize the described aspect.

MAY. An aspect or requirement prioritized with this term is to be understood as "...truly optional... " [68].

The priorities assigned to the requirements in this document reflect an initial assessment of the consortium based on their domain expertise, particularly from the technical partners and the use case leaders. During the requirements refinement process planned later in the project, it is aimed to revalidate the prioritisation in a use-case-driven approach. Hence, as (1) the SPATIAL use case partners are representatives of the identified stakeholders and (2) their use cases are instances of the AI-based systems we aimed to collect design guidelines and recommendations for, we expect that the consortium's assigned priorities will be also meaningful and relevant for AI-based systems developed outside of SPATIAL.

In this context, we want to emphasise that the priorities defined in this document are explicitly not binding for the prototypes of the use cases or the Explanatory AI Platform to be developed in SPATIAL. Instead, the assigned priorities reflect the consortium's assessment of aspects that

[4] It is important to notice that is intended and required to write the terms in the uppercase format.

production-ready AI-based systems from TLR8 onwards should strongly consider. Since none of the SPATIAL prototypes will reach this TRL level, we distinguish the requirements between those relevant to the activities of the SPATIAL use cases and Explanatory AI platform, and those generally valid for designing and developing trustworthy AI-based systems ready for operational environments. In the following section, we explain this classification in detail.

3.1.3 RELEVANCE OF THE COLLECTED REQUIREMENTS

Although this document aims to provide a universally valid catalogue of requirements expressing guidelines for designing AI-based systems developed in the context of the four domains mentioned in Section 1[5], some of the requirements formulated in this document are still highly relevant for the technical activities of the SPATIAL project (see Section 1.2 and in Figure 1). Therefore, in order to be able to distinguish more precisely between requirements universally relevant for designing AI-based systems and those also relevant to the technical activities in the SPATIAL project, we have assigned a relevance to each identified requirement in Appendix A (see column "*Relevant for*" in Appendix A). These relevance scores are defined as described below. In this context, it is essential to note that a single requirement can have multiple relevance scores assigned to it, meaning it can be relevant for the use cases, the platform, and in general.

Relevant in General: This label was assigned to requirements that should be generally considered in the design, specification, and development of AI-based systems. Although they represent relevant recommendations, these requirements do not need to be considered in the context of SPATIAL, as they are either out of scope, not relevant for the use cases in their current TLR level, or refer to operational aspects in production environments.

Relevant for SPATIAL Use Cases: As concrete representatives of AI-based systems, some of the gathered requirements are also highly relevant for at least one of the four SPATIAL use cases. Hence, we assigned the relevance *"Relevant for SPATIAL Use Cases"* to these kinds of requirements in Appendix A. These should be considered in the pilots during the design and development of the respective use cases. In this context, it is important to remember that the four SPATIAL use cases refer to the individual SPATIAL key pillars (i.e. privacy, accountability, resilience) and only reflect certain aspects of these. Hence, none of the use cases will meet all requirements assigned with this relevance.

Relevant for SPATIAL Platform Components: The catalogue provided in this document also contains concrete aspects that are highly relevant for shaping the design and specification of the *Explanatory AI Platform* developed in WP3 (see Section 4). Therefore, we have assigned the relevance *"Relevant for SPATIAL Platform Components*" to these kinds of requirements in Appendix A. These requirements represent recommendations or accountability needs that the

[5] The aim of D1.1 is to gather recommendations and general design guidelines for AI-based systems in the technical contexts of Mobile Edge Systems (e.g. 5G Services), Cybersecurity Applications and Analytics, IoT, and eHealth in the form of requirements.

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

platform should address to be beneficial for the four SPATIAL use cases or general AI-based systems. Based on these kinds of requirements, it should be possible to identify required platform functionalities and shape corresponding services (see Section 4).

Traceability, tracking, and evaluating the fulfilment of relevant requirements

This document aims to establish a comprehensive catalogue of aspects and general design principles to consider when integrating and utilizing AI in modern system architectures. Nevertheless, as highlighted in the recent paragraph, some of these requirements are relevant to the project's technical activities. For this reason, we introduced the differentiation between the relevance of the requirements as described above. According to the assigned relevance, these design goals or recommendations[6] should then be considered in the diverse technical activities of the SPATIAL project. In order to track, which of the requirements are taken into account and fulfilled in the project, we plan to adopt two traceability approaches. On the one hand, we aim to refer in the corresponding technical deliverables of WP3 and WP5 with respect to which individual requirements have been considered in the technical activities. On the other hand, we plan to provide a separate document[7] explicitly summarising which of the requirements was considered and how. This document will also indicate how and by which technical activity (e.g. use case or Explanatory AI platform) the respective requirement has been addressed.

Furthermore, deliverable D1.3 has the goal of refining these abstract design principles to more specific system and design requirements. This refinement will be based on the experiences gathered during the prototyping and the belonging system requirements are continuously updated during the design and implementation process. Hence, traceability will be established also between the requirements from D1.3 (being a refinement of D1.1) and the way those are addressed within the final prototype. The overall traceability relations will be summarized in a separate document – as mentioned above – resembling a traceability matrix between requirements and evidence within the prototype and the executed experiments/tests.

3.2 SCOPE AND LIMITATIONS: THE SPATIAL USE CASES

In the following section, we will present the four practical use cases realized in the SPATIAL project. In addition, we will discuss the challenges and problems of the use cases that will be addressed in SPATIAL. In this context, the presented use cases are intended to verify, validate, and showcase the contributions of the SPATIAL project first-hand in the scope of industrial environments and applications. Thereby, it is important to notice that the practical use cases reflect the main topics considered in SPATIAL: Mobile Edge Systems (e.g., 5G Services), Cybersecurity Applications and Analytics, Internet of Things, and eHealth.

[6] Please note that most of the requirements present design guidelines and recommendations for the design and development of AI-based systems.

[7] This document will most likely be provided in the form of an Excel sheet.

3.2.1 USE CASE 1: PRIVACY-PRESERVING AI ON THE EDGE AND BEYOND

Our first use case provided by Telefonica Investigacion Y Desarrollo SA (TID) envisions an environment where multiple machine learning applications are using personal data at a large scale. While such large-scale machine learning applications have been predominantly centralized (i.e. all data of users/clients/devices need to be uploaded to cloud platforms for learning), the recent advances in Federated Learning (FL) allows to build ML models in a decentralized fashion close to users' data, without the need to collect and process them centrally. However, individual machine learning applications still have to separately employ the mechanisms for privacy preserving [69] and distributed model development [70].

In this context, Use Case 1 aims to demonstrate an edge-based federated learning platform on top of the telecommunication infrastructure, which enables the applications to collaboratively build machine learning models, while protecting users' privacy and abstracting the operation of model training and transfer. The platform enables different applications to build better models thanks to larger and richer datasets and supports developers and data scientists to easily and quickly deploy federated learning solutions, fostering a large adoption of the federated learning paradigm.

We provide a few example scenarios below:

Example scenario 1: A unique federated learning modelling per individual application for an existing machine learning problem could be created without the need of developing the algorithms. Traditionally, a machine learning modelling is requested uniquely per application aiming to solve a specific, existing ML problem: e.g. a streaming music application (e.g. Spotify) that wants to model its users' music preferences to provide better recommendations.

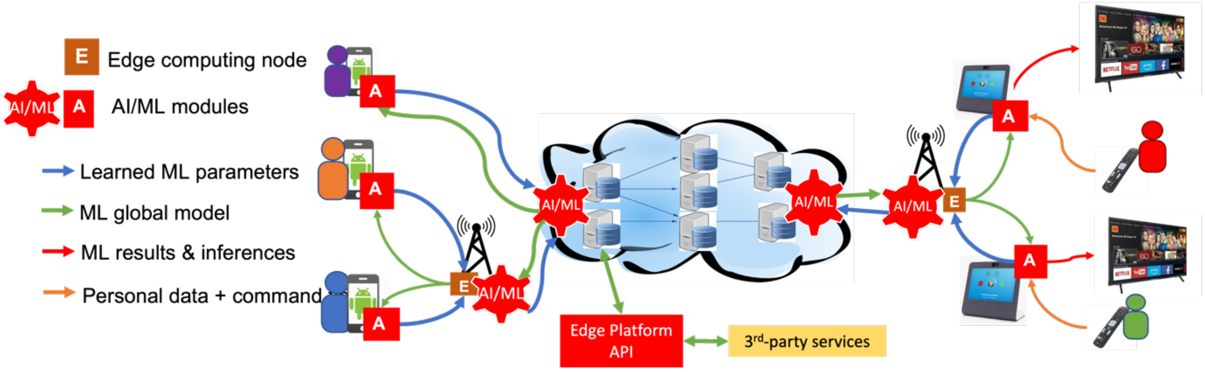

FIGURE 4: ARCHITECTURE AND DATA FLOW OF THE PRIVACY-PRESERVING EDGE AI TRAINING PLATFORM

Example scenario 2: A unique federated learning model can be trained in a joint fashion between two or more collaborative applications for an existing machine learning problem. That is, a group of applications interested in collaborating to build a shared machine learning model that solves an existing problem, identical and useful for each application, but on more, shared and

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

homogeneous data. For example, Instagram, Messenger, and Facebook (owned by the same company) may want to build a joint machine learning model for better image recognition, on images of similar quality and scope, but coming from each application's local repository.

Example scenario 3: A unique federated learning model can be trained in a joint fashion between two or more collaborative applications, as in case (2), but for a novel ML problem. For example, an application for planning transportations (e.g., Uber, GMaps, or Citymapper) may want to model your music preference considering the mode of transportation (e.g., bicycle, bus, car, etc.).

We expect the following requirements in realizing our platform.

Permission management across applications and services: Mobile and IoT systems provide mechanisms to grant application and services access to data such as mobile sensors, location, contacts, or calendar. Such access is typically given at a very coarse granularity (e.g. all-or-nothing), and can be unrestricted or, more recently, granted per application. On top of these traditional permissions, our platform has to provide mechanisms to specify permissions across applications and services to share data and models among them. Further, it has to provide security mechanisms to guarantee these permissions are respected.

Privacy-preserving schemes: In the deployment scenarios and use cases of our platform, multiple applications and services can be involved in the federated learning execution. In order to guarantee the privacy of customers' data, it is critical to leverage privacy-preserving mechanisms in the construction of federated learning models. In our platform, we plan to leverage Differential Privacy (DP) to provide further privacy guarantees to participating clients. DP noise can be introduced at different stages of the federated learning system: in the data source at the client side, also known as local-DP [71], at the central server side [69] while building the global model, or at an intermediate stage such as edge computing nodes [72] or base stations [73], or with hybrid methods and hierarchical methods [74] [75]. However, introducing DP noise in the ML pipeline reduces model utility [76] as it affects convergence rate of the federated learning-trained model.

Exchange model across a (hierarchical) network with federated learning: As depicted in Figure 4, Federated Learning as a Service (FLaaS) can build models in a hierarchical fashion across different network layers: end-user device, Internet Service Provider (ISP) edge nodes, or the central server. Recent works considered the hierarchical federated learning cases, where multiple network stages are involved in the training process [77]. Such efforts showed convergence and accuracy can be improved with proper design under such settings. The Telefonica platform will build on these works to realize its hierarchical use cases.

Training convergence and performance: As mentioned earlier, the usage of differential privacy in multi-stage training and the hierarchical federated learning approach impact the convergence and performance of the models. However, in the Telefonica platform, the possibility of building cross-application models introduces another dimension, potentially impacting model convergence and performance.

Platform usability: Every service platform should enable a customer to use its services with limited overhead and knowledge of the underlying technology. On the one hand, existing commercial Machine-Learning-as-a-Service platforms (e.g. AWS, Google Cloud or Azure) provide users with Application Programming Interfaces (APIs) and Graphical User Interfaces (GUIs) to configure and use the services in cloud environments. However, these APIs are not designed to deal with cross-application model building, nor tailored for federated learning services. On the other hand, existing federated learning libraries (e.g. OpenMined [70]) are still in prototype phase and cannot support a service model, and do not provide GUIs, or high-level service APIs. They also do not support cross-application modelling as our platform does. The Telefonica platform builds on these existing works and should provide high-level APIs to support model building across applications on the same device and across the network, and software libraries or Software Development Kits (SDKs) for application developers to include the service in their apps and devices.

3.2.2 USE CASE 2: IMPROVING EXPLAINABILITY, RESILIENCE, AND PERFORMANCE OF CYBERSECURITY ANALYSIS OF 4G/5G/IOT NETWORKS

4G/5G and IoT networks hold the promise of delivering ultra-low latency, ultra-high throughput, ultra-high reliability, ultra-low energy usage, and massive connectivity. Technical security measures toward 4G/5G/IoT networks are quickly embracing a variety of machine learning algorithms as an effective approach to empower intelligent, adaptive and autonomous security management, allowing to tackle the growing complexities of the network. Indeed, AI has the potential of recognizing abnormal patterns from a large set of time-varying multi-dimensional data, and delivering faster and accurate decisions [78]. However, the adoption of AI methods in IoT and future mobile networks is still in its infancy. Thus, research efforts need to take into consideration diverse aspects of AI solutions and practical implementation issues to support both users and developers in effectively auditing the code and data of safety-critical systems. However, AI-based systems have three major issues in the 4G/5G/IoT domain as follows:

Lack of real-world datasets. AI models, such as supervised ML, require large amounts of data with correct labels, so the quality of data has a great impact on the advancement of modern AI research. However, diversified real-world datasets in 4G/5G/IoT networks are not readily available due to privacy issues that need to be followed by all telecom operators.

Lack of explainability. Currently, AI schemes applied in security solutions focus mainly on accuracy and performance (e.g. precision, recall, resource utilization) and do not readily offer an explanation of why a particular output is obtained. Explanation of a decision taken often

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

becomes a critical requirement for the 5G network, especially because many critical services depend on the 5G infrastructure.

Lack of resilience against adversarial attacks. ML models are vulnerable to adversarial attacks [79] where adversarial inputs are small carefully crafted perturbations in the test data built for fooling the underlying ML models into taking the wrong decisions. Robustness against adversarial attacks is a challenging problem as there does not exist a solution that ensures complete protection against this kind of attacks.

3.2.2.1 Proposal

In the context of the Use Case 2 lead by Montimage (MI), we aim to tackle the above challenges by: (1) producing real-world datasets for AI training, especially for 4G/5G and encrypted network traffic thanks to our open source 5Greplay tool [80]; (2) enabling explainability features of existing AI algorithms in our different AI-based systems, such as MMT-Probe[8] for anomaly detection and MMT-RCA for Root Cause Analysis (RCA); and, (3) considering the security threats that emerge from the rapid adoption of AI algorithms in 4G/5G/IoT networks. Existing challenges of AI in 4G/5G/IoT networks motivate us to make the ML models more *accurate*, *explainable*, and *robust* before they are integrated into complex systems. To do so, we provide a *real* 4G/5G/IoT testbed that allows us to evaluate various AI techniques related to the explainability, resiliency, and distribution of AI models. This will be done by assessing the techniques used today by Montimage in its Montimage Monitoring Tool (MMT) security monitoring framework [81] for performing cybersecurity analysis and protection of 4G/5G/IoT networks, encrypted traffic analysis and RCA, in the H2020 SPATIAL project.

3.2.2.2 Overview of Our Testbed

Figure 5 shows the overall architecture of the real 4G/5G/IoT testbed leveraged in Use Case 2. This pilot corresponds to a 4G/5G/IoT solution consisting of an gNodeB based on a Software-Defined Radio, a portable 5G Core solution, and the MMT security monitoring framework. The framework is based on distributed and extensible MMT-Probes that analyse (encrypted) network traffic from both the mobiles and IoT devices. The MMT-Operator is a Web application that allows further analysis and acts as a decision point for determining the causes of breaches and triggering corresponding countermeasures. Our AI-based tools, including MMT-Probe for cybersecurity analysis of (encrypted) network traffic and MMT-RCA for Root Cause Analysis, will later be evaluated with respect to the improved performance, transparency, and precision they bring to the security analysis and algorithms.

[8] MMT stands for Montimage Monitoring Tool.

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

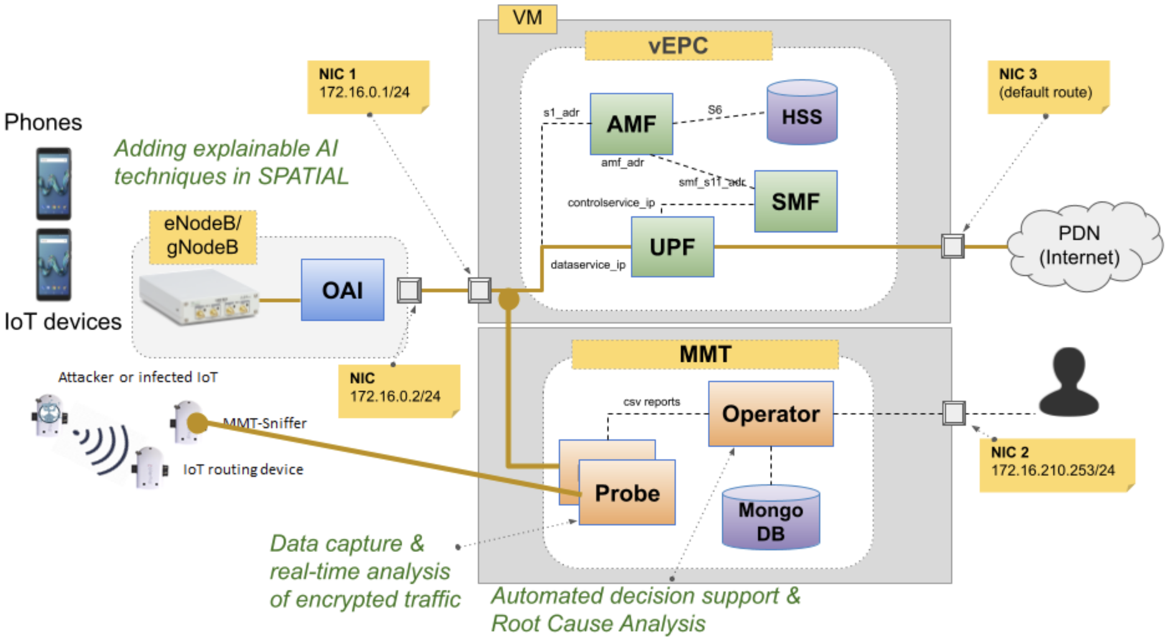

FIGURE 5: ARCHITECTURE OF THE 4G/5G/IOT TESTBED IN USE CASE 2

IoT testbed. Our IoT testbed [82] includes a set of equipment forming an IoT IPv6 over Low-Power Wireless Personal Area Network (6LoWPAN): several Zolertia REMotes, one Raspberry Pi and accessories. A border router mote acts as the gateway collecting sensed data from the other motes, and forwards the reports via the Universal Serial Bus (USB) line to the server deployed in the Raspberry Pi. The IoT network consists of normal clients reporting sensed data every 10 seconds, and one (or several) attacker(s) behaving interchangeably in one of three modes: (1) *normal mode* reporting data every 10 seconds, (2) *denial of service attack mode* reporting data 100 times faster (10 messages/second), and (3) *dead mode* not reporting data at all. An IoT device, MMT-Sniffer, performs sniffing tasks by capturing network traffic and piping it via the USB line to the Linux-based machine where MMT-Probe is deployed to analyse the traffic and extract the metrics for our MMT-RCA. A Raspberry Pi feeds the motes in terms of batteries, hosting the server dealing with the sensed data and receiving the sniffed traffic which is then analysed by the MMT-Probe.

4G/5G testbed. EPC-in-a-Box [83] platform represents a 4G/5G network core commercialized by Montimage and Cumucore. It is a ready-to-use appliance allowing the creation of a full end-to-end 4G/5G network in 5 minutes. It can be used not only for testing but also to create a small-scale mobile network in order to provide mobile connection in industry. It basically consists of 3 main building blocks: RAN, Evolved Packet Core (EPC) or 5G Core, and MMT. Once deployed, the testbed platform creates a 4G/5G network allowing commercial off the shelf UEs to connect. After being successfully attached, the UEs have access to the Internet Protocol (IP) services in Public Data Networks (PDNs) or from the Internet such as browsers, Web applications, Voice over IP (VoIP) video calls, etc. All the traffic between the RAN and the EPC is captured and analysed in real-time by MMT-Probe to ensure that the defined security properties are respected. MMT-Operator supports automated decision and reaction in the case an anomaly is detected.

To summarize, feature selection, similarity learning, and Bayesian networks are currently employed in our AI-based modules in the 4G/5G/IoT context, but these techniques need to be improved to make them more resilient to attacks by considering adversarial attacks, transparency and explainability. To improve the explainability of AI models, we will apply well-known XAI techniques, such as LIME, SHAP, and new ones developed by the SPATIAL project partners into our AI-based modules. Furthermore, in view of increasing the resilience to AI threats, we will consider and implement different evasion and poisoning attacks against ML models used in cybersecurity detection and management.

3.2.3 USE CASE 3: ACCOUNTABLE AI IN EMERGENCY ECALL SYSTEM

Emergency calls (eCalls) and their underlying emergency communication system are an extremely critical matter for the aging society of Europe. In order to inform about and respond to emergencies quickly and effectively, an efficient, modern, and reliable emergency communication system is of utmost importance. For this reason, a significant amount of research effort has been invested in the evolvement of existing emergency communication systems in recent years. A recurring concept in this context is that of Next Generation 112 (NG112) emergency communication systems. These systems represent an evolution of traditional phone-based emergency communication systems. Typically, NG112 emergency communication systems are based on IP networks in which emergency calls are performed via VoIP calls. Compared to traditional emergency communication systems, this offers several advantages. As an example, in addition to pure voice information, VoIP emergency communication also allows the transmission of rich-media information. For instance, the initiated emergency call can also take the form of video calls. Furthermore, the IP-based infrastructure also enables to share additional information, e.g. sensor data of a patient, directly with the emergency call center during an emergency call. The ubiquitous integration of IoT devices and eHealth sensors is continuously gaining momentum in the healthcare sector due to their availability, accessibility, and cost-effectiveness [84] [85]. Wearables such as smartwatches, smart T-shirts, or chest straps enable collecting and monitoring diverse and relevant health data such as a patient's blood pressure, heart rate, oxygenation, blood glucose levels, or body temperature [84] [86]. This data can also be extremely relevant and helpful in the case of emergencies. Since NG112 emergency calls enable the transmission of such eHealth sensor data, emergency call center personnel can better assess the emergency situation and initiate more appropriate countermeasures. Furthermore, this additional information also enables an improved estimation of the required medical effort and a more targeted allocation of medical personnel. In addition, the eHealth sensor data and NG112 infrastructure enable the emergency doctor in advance to better assess the emergency, prepare necessary life-saving measures, or instruct first aiders. Besides, the optional sensor data provided to the emergency call centers enables them to identify hoax calls, which saves medical resources.

3.2.3.1 The EMYNOS Project and the EMYtest Testbed

In the context of the EU-funded H2020 project EMYNOS[9] (nExt generation eMergencyY commuNicatiOnS), Fraunhofer FOKUS was involved in the specification and development of a prototype for an NG112 emergency communication system. The EMYNOS project aimed to design, define, and implement a platform for IP-based bidirectional emergency communication that overcomes the limitations of traditional systems, e.g. in the transmission of caller location, the forwarding of emergency calls, or the integration of sensor data and IoT architectures. An important focus was set on the requirements of people with disabilities by using eHealth sensors to implement appropriate monitoring and the resulting automatic emergency calls. The high-level architecture of the EMYNOS platform has been discussed and specified in detail by Rebahi et al. [87]. In addition, a testbed called EMYtest that follows this specification has been developed for evaluation and demonstration purposes as part of the EMYNOS project. The EMYtest testbed is shown in Figure 6 and described by Kumar Subudhi et al. [88] and Barakat et al. [89]. As can be observed, EMYNOS differentiates between the Call Origin Side, the Network Side, and the Public-Safety Answering Points (PSAP) Side.

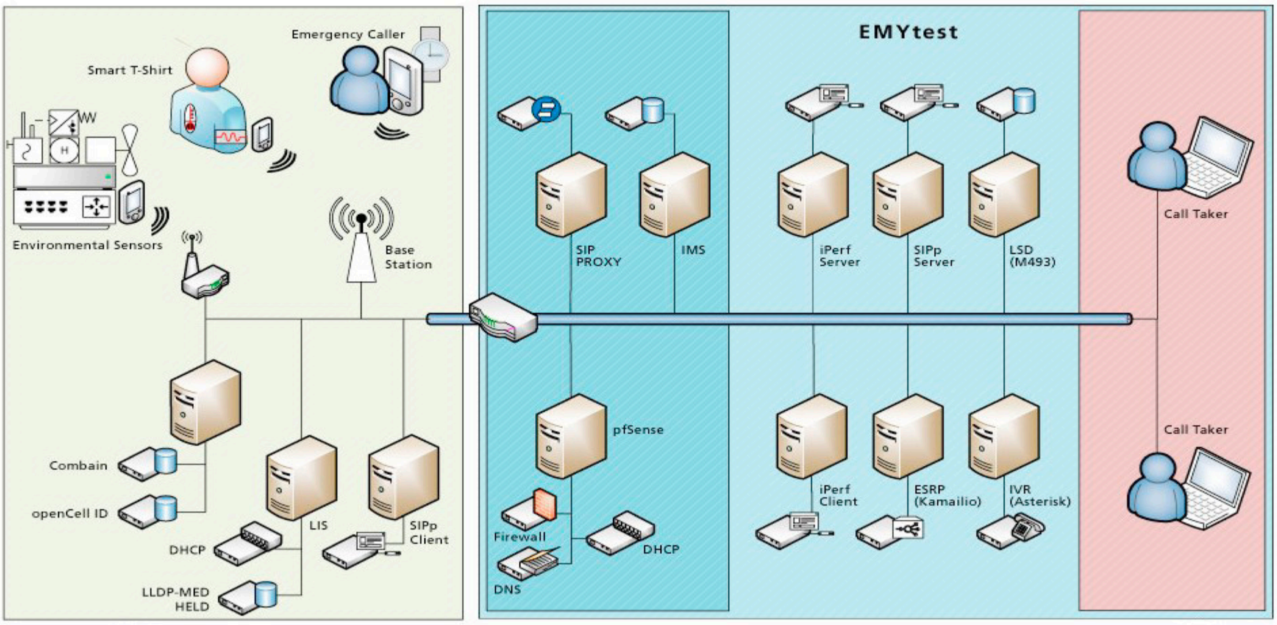

Figure 6: Prototype and testbed of the EMYNOS platform deployed in the data center of Fraunhofer FOKUS [88]

[9] EMYNOS: https://www.emynos.eu/, as of date 22.02.2022

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

Call Origin Side

As the name already implies, the Call Origin Side - illustrated in green in Figure 6 - initiates an emergency call. In EMYNOS, the emergency call can be initiated from various VoIP-enabled devices (e.g. smartphones, smart TVs, smartwatches, or desktop computers) and can be triggered manually or automatically. However, in the EMYtest testbed, all emergency calls are currently only carried out via an Android-based smartphone. In the EMYtest testbed, the Call Origin Side is also equipped with multiple eHealth sensors (e.g. smart T-shirts) that send their measurement data to the smartphone. This data is then processed and analysed (on the smartphone) and, if necessary, attached to the VoIP emergency call as an optional payload. Furthermore, the Call Origin Side is also capable to resolve and attach the location of the emergency caller by leveraging a Location Information Server (LIS).

PSAP Side

The counterpart to the Call Origin Side is the PSAP Side shown in red in Figure 6. Emergency calls received at the Public-Safety Answering Points are routed through this switching facility to an available and free emergency call recipient (Call Taker). The Call Taker receives the emergency call, interacts with the caller, and initiates appropriate countermeasures. In this process, the Call Taker also uses a Linphone-based software to perform the VoIP-based eCall.

Network Side

The Network Side - depicted in blue in Figure 6 - is responsible for establishing the VoIP-based emergency call between the Call Origin Side and the PSAP Side. For this purpose, the Network Side determines the most appropriate PSAP based on the emergency caller's location and then routes the eCall to it. In this context, call identification, call classification, and the location to PSAP mapping play an essential role [3]. The emergency caller's device first sends a message to the Emergency Services Routing Proxy (ESRP). The ESPR is then in charge of identifying and classifying incoming calls, determining the responsible PSAP by querying a location-to-service server, and finally routing the eCall to the determined PSAP.

3.2.3.2 The EMYNOS Extension envisioned in SPATIAL

At its current version, the EMYNOS NG112 platform provides valuable emergency communication functionality. However, in order to deliver an even higher quality service, the platform could be further extended with additional mechanisms and tools. Therefore, our primary goal in the SPATIAL Use Case 3 is to fully automate the process of emergency detection and the ensuing triggering of an emergency call. This could be achieved with the help of state-of-the-art AI-based approaches, which are well equipped to detect anomalous behaviour in an automated manner. In simple terms, we aim to extend the current EMYNOS NG 112 platform by implementing an automated, effective, accountable, and privacy-preserving AI-based system that operates on heterogenous eHealth data and can accurately recognize emergency situations and initiate eCalls when these situations occur. Such a system could provide critical help, especially in high-risk, life-threatening, crisis situations where the delay between the emergency occurrence and the medical intervention must be minimized (e.g. an elderly person falls over). In order to minimize this delay and to reduce the number of health-related emergencies that are not handled on time, there are multiple important challenges that the extended EMYNOS platform has to address.

First, the extended EMYNOS system needs to **provide the necessary hardware and software mechanisms** that support the complete communication flow between the involved parties (e.g. injured person and eCall receiver). In terms of hardware, we can leverage the rapid integration of a wide variety of IoT devices and health sensors such as Wearables and their emerged integration into body area networks (BANs) and wireless body sensor networks (WBSNs). However, one crucial challenge remains the sensitive nature of data generated by such health sensors. The **collected and processed healthcare data is to be classified as highly sensitive personal data, which should be protected at all costs**. In addition, given that the extended EMYNOS system would incorporate AI-based emergency detection mechanisms based on this data, it must provide privacy-preservation, accountability, and security guarantees. In this context, the **employed AI model must also be robust against data poisoning and adversarial attacks**. These manifold aspects are not only requirements that align with the goals of the SPATIAL project but also properties that can support the trust establishment between the end-user and the platform. Therefore, the resilience and accountability measures, the privacy-preserving methods, and the guidelines for integrating secure and trustworthy AI envisioned in SPATIAL will be of great importance for this use case.

Another critical challenge regarding the EMYNOS extension is related to the **collection and utilization of the required eHealth data**. As already mentioned above, we aim to use healthcare information such as blood pressure, heart rate, oxygenation, blood glucose levels, or body temperature collected from various IoT devices and eHealth sensors to detect emergency situations and automatically initiate emergency calls. Due to the heterogeneous nature of this data originating from various sensors, it will be challenging to **collect, clean, and unify a data set that is representative, fair, balanced, and unbiased.** However, this is a key requirement for successfully applying automated emergency detection. The data collected must be representative of all genders and age groups so that the AI-based detection of emergencies can be made regardless of age and gender. It is also important to ensure that the data also reflects historical health conditions and current fitness levels so that these aspects can also be considered in emergency detection. Finally, the labelling of the gathered health data can be costly since it is possible that medical experts must be involved in this process.

In addition to the above aspects, the **performance of the employed AI model** is of utmost importance in the context of the safety-critical domain relating to this use case. Emergency calls are a very critical aspect and need to be initiated at the right moment. At the same time, it must be avoided to overload the emergency call centers in the backend. Therefore, on the one hand,

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

we need to develop an effective and accountable AI model that detects emergencies at a rate as high as possible, in order to provide help in life-threatening emergency situations. This could be achieved by increasing the model's rate of true alarms. On the other hand, it must be guaranteed that the AI model does not initiate too many unnecessary or false emergency calls so that the emergency call takers are not overloaded and blocked for real emergencies. In contrast to the former goal, this could be achieved by decreasing the model's rate of false alarms. Therefore, effective measures and a trade-off between these conflicting goals must be found, in order to create a good-performing model that reliably detects emergencies but does not overload the emergency call centers in the backend.

Finally, the **utilization of XAI methods** on the platform would be not only beneficial for the end-users, but even more so for the recipient of the emergency eCalls. More specifically, if the call is accompanied by an explanation about the model's decision, the recipient of the call could potentially assess the degree, to which the particular case is life-threatening. The explanation could contain information that indicates in advance how the medical staff could prepare for handling the emergency situation. Given the highly sensitive nature and high resource demand generated by eCalls, XAI explanations could also help recognize false emergency alarms that could be ignored in favour of real emergencies. Therefore, using XAI methods in the course of the EMYNOS system extension is a pre-requisite requirement aiming to ensure that the platform utilizes resilient, accountable, privacy-preserving, and secure AI methods. Once again, this aligns with the goals of the SPATIAL project and its expected methods and guidelines for developing explainable, transparent, and trustworthy AI, which will once again be highly useful.

3.2.4 USE CASE 4: RESILIENT CYBERSECURITY ANALYTICS

Many cybersecurity functions are now fully, or partly, driven by automated data analysis enabled by artificial intelligence and machine learning methods. ML classifiers are used to detect malicious contents such as malware, malicious documents, phishing websites, spam emails, etc. ML-based anomaly detectors leverage their ability to quickly analyse and find anomalies in large amounts of network traffic to efficiently detect intrusions in networks. In parallel to automated decision making, machine learning is also used to support human decision in security-related issues. For instance, clustering is used to group and synthesize security alerts and attack reports in managed detection and response (D&R) systems, improving their presentation to human analysts. Machine learning also supports the manual investigation and forensic analysis of attacks by human security operators in security operation centers (SOCs), prioritizing and enriching existing security events with additional information.

The diversity of cybersecurity applications is associated with different requirements for ML models that are employed in them. The main difference for requirements is between ML used to support human analysis and decision, and ML models used for automated decision making without human supervision. The involvement of humans or not in the final decision-making drives very different requirements especially on the reliability of the decision that must be

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

provided by ML-based systems. In this project, we will mostly focus on the use case of resilient cybersecurity analysis based on machine learning for automated decision making and without human supervision. This choice was made to follow the current trend that promotes the usage of ML to make systems more autonomous and less reliant on humans. This trend is the most prominent when it comes to applying AI and ML in the core topics we tackle in this project: Mobile Edge Systems, Cybersecurity Applications and Analytics, IoT, and eHealth. Two representative applications are selected and will be studied for the resilient cybersecurity analysis Use Case 4 organized by F-Secure (FSC).

Our first cybersecurity analysis application relates to ML used for automated malicious content detection. Our case study is an ML-based system that differentiates malicious from benign documents, e.g., Microsoft Word documents. Features defined by security experts are automatically extracted from documents that are unknown in nature. These features are input to an ML classification model that renders a decision whether the document is malicious or not. The ML model used here is a supervised binary classifier trained using documents that are known to be malicious or benign. This malicious document detector is a typical example of malicious content detection based on machine learning. An unknown content, e.g. software program, document, website, email, is transformed into numerical features that are input to a binary classifier, which automatically decides whether it is malicious or not.

Our second cybersecurity analysis application relates to ML used for modelling system and user behaviours. Such models are typically used for automatically detecting anomalies that deviate from the learned "normal" behaviour. They are also used for automatically identifying unknown systems and users, e.g., to apply more effective security measures that are specific to the identified system or user. Our case study is an ML-based system that identifies the type of an unknown host by analysing the processes it launches, the domains it accesses, and the files it opens. The types of identified host can be for instance, server, technical user, non-technical user, etc. This host-type identification system uses an ML classification model that predicts as many classes as there are types of hosts to identify. It is trained in a supervised manner using features representing the processes launched, domains and files accessed by hosts, from which the types are known. This host-type identification system is a typical example of behaviour model based on machine learning for cybersecurity. Different security policies are applied to different types of hosts, to detect host compromise more effectively.

In both our case studies, we focus on the usage of automated decision making for cybersecurity applications. This requires optimizing the accuracy of ML models in a specific manner and maximizing the average accuracy of the system is usually not the end goal in these use cases. In contrast, we often aim to minimize the false detections (or false positives) at the price of missing the detection of some attacks or malicious content. The detection of malicious contents automatically prevents their access to users and any false detection negatively impacts usability. On the other hand, ML-based detection systems for automated decision making are usually not stand-alone, and they are completed with more traditional rule-based and/or signature-based detection systems that can cope with missed detections. Thus, the negative impact of missed detections on security is lower than the negative impact of false detections on usability. To enable a customized optimization of the accuracy, the ML models used in these cybersecurity applications need to provide soft prediction scores rather than hard decisions. The soft prediction scores can be used with a tuneable threshold to trade-off false detections and missed detections as required. Several thresholds can also be used such that we could choose not to provide a decision if the prediction score is too uncertain. Thus, when studying this use case, we will primarily focus on ML models that can provide soft prediction scores.

ML models used in cybersecurity analysis are often conceptually simple and supervised classification methods like Logistic Regression, SVM, Random Forest or Gradient Boosting are examples of the most used type of models. Deep Neural Networks (DNNs) are seldom used because their ability to process natural data like images, sound, text, etc. using raw features (e.g., pixel value) is not required in cybersecurity analysis. Features used for cybersecurity analysis are most often defined and engineered by security experts using their expert domain knowledge. This is the case for features used in our malicious document detector and host-type identification system. Considering the refinement of these features, simple ML models are often very effective in classification tasks and that is why they are the most used ones. They are cheaper to train, easier to tune and their decision is more understandable. Consequently, this use case will mostly focus on the study of requirements for ML-based systems that use simple ML classifiers and most likely non-DNN models.

Our two proposed case studies rely on supervised ML classifiers that are trained using labelled data. Hence, we will focus on studying requirements for ML-based systems of resilient cybersecurity analysis that use supervised ML models. The performance and the accuracy of supervised ML models is tightly linked to the quality, representativity and integrity of their training data. Consequently, model requirements are highly dependent on data requirements, and it could be difficult to define them separately. Similarly, the fulfilment of model and data requirements could be challenging to evaluate separately, and it will likely require a unified approach to evaluate them together.

ML models used in cybersecurity analysis are often part of a complex and hybrid automated decision system. As pointed out before, the decision of ML models is often combined with rule-based and/or signature-based systems. The requirements on the different components involved into making a final decision are dependent on how individual decisions are combined. Similarly, these requirements will be dependent on the way the system is deployed and on the dependability between these different components. Thus, having defined global requirements, the local requirements for components of the system, like the ML model, may be difficult to single out in complex systems used for cybersecurity analysis.

In the cybersecurity domain, the existence of attacks and attackers trying to compromise deployed defences is a given. Attackers attempt to circumvent any protection that represents a barrier between them and their attack target. When ML models are used to protect information systems, to detect and prevent attacks; it is certain that attackers will attempt to compromise their integrity, such that they will make incorrect decision, e.g., failing to detect attacks. Because of the prevalent security threats, the security requirements will be at the core of the resilient cybersecurity analysis use case. A particular attention will be given to enforce the resilience to attacks that specifically target ML-based systems, which are known as *adversarial machine learning attacks* [91].

To summarize the scope and limitations of the resilient cybersecurity analysis use case: it will be focused on two representative cybersecurity applications that use ML for automated decision making. The study of requirements will be focused on systems using simple supervised ML classifiers that can provide fine-grained prediction scores. A challenge will lie in the complexity of identifying requirements for individual components of these ML-based systems due to their interdependency (e.g., model and data) and their complexity (many components involved in final decision-making). Finally, Use Case 4 will have a prominent focus on the security requirements since it is paramount for cybersecurity applications and more important than for other core topics: Mobile Edge Systems (e.g., 5G Services), IoT, and eHealth.

3.3 USER AND STAKEHOLDER IDENTIFICATION

In the following section, we will identify and discuss stakeholders that are most relevant for the activities and objectives of the SPATIAL project. Precisely, we will discuss the stakeholder *End Users*, *Developers*, *Auditors*, *Testers*, and *System Operators*. Figure 7 summarizes these stakeholders and visualizes their level of power and interest regarding the SPATIAL project in a stakeholder matrix.

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

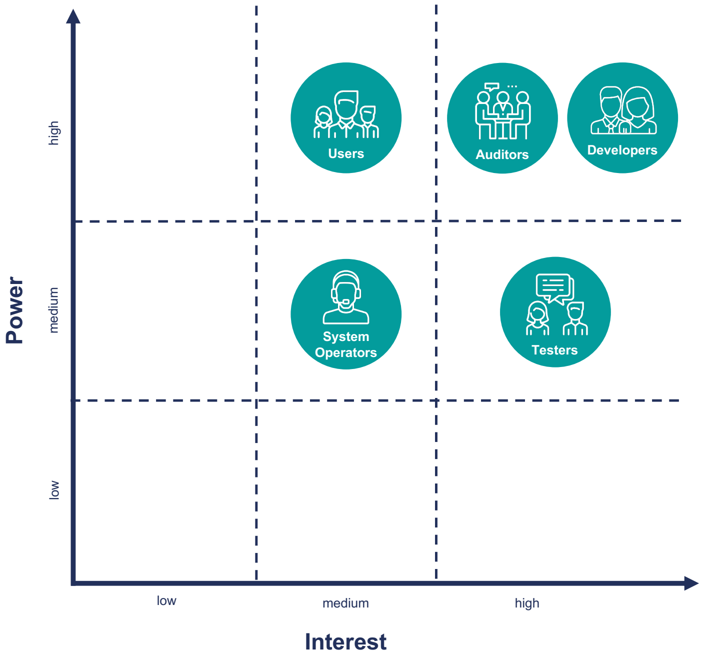

FIGURE 7: STAKEHOLDER MATRIX VISUALIZING THE LEVEL OF POWER AND INTEREST OF THE IDENTIFIED STAKEHOLDERS

3.3.1 DEVELOPERS

Developers are highly relevant stakeholders of the SPATIAL project, since they are responsible for the delivery of high quality, secure, AI-based solutions. The developer community in this case can be divided into two subgroups: ordinary software developers and developers of artificial intelligence solutions. While the SPATIAL project will largely aid developers of AI solutions, as it's a group that requires new directions, the deliverables of WP2 and WP3, will benefit both groups. In WP2, the partners will establish a design approach for incorporating resilience measures into AI algorithms, building on the mechanisms developed in WP3 for enhancing the robustness of AI systems trained and deployed in an uncontrolled environment. This will enable developers and providers of AI-based technologies and services to defend against significant kinds of attacks, assess their systems' reliability limitations, and mitigate residual risks. When developers and service providers have metrics for determining what attackers can accomplish and with what resources and capabilities, they will be in a better position to make optimal trade-offs and to provide relevant evidence of their systems' resilience properties to customers, standardisation bodies, and regulators.

By equipping developers with the tools and knowledge necessary to create more transparent processes and explainable AI, the SPATIAL project will begin to alter the culture of AI development at an early stage - that of aspiring engineers and computer scientists. Furthermore, SPATIAL will offer realistic guidance for software developers on how to modify AI systems, ensuring a streamlined application of trustworthy AI in cybersecurity solutions. Additionally, it will facilitate clear communication between end users, testers, auditors, and system operators regarding the use of AI in cybersecurity.

Level of Interest: high

Level of Power: high

3.3.2 END USERS

Another main group of identified stakeholders interested in the activities and outcomes of the SPATIAL project are the **End Users**. This group is highly diverse and represented by many manifold actors. It compromises, among others, policymakers, business partners, but also the average consumer of AI-based products, systems, and networks. In general, end users of AI-based systems, networks, and security solutions expect the correct functioning of these systems according to their promised specifications, comfortable usability, and a clear and easy-to-understand interface - including simple explanations (e.g. through visualizations). Besides, they want the used systems to be secure and resilient at the highest possible level. Furthermore, users want systems to be accountable and trustworthy, especially in safety-critical applications. In this context, users also want to easily understand and comprehend the decision-making of AI applications and their observed behaviour. This is fundamental to increase the trust and acceptance of AI-based applications and systems. As a result, explainability and transparency are relevant aspects for them. Therefore, we conclude that the stakeholder group of end user is highly interested in the activities and potential outcomes of the SPATIAL project. Precisely, the envisioned accountability and resilience metrics and the planned system solutions, platforms, and standards that are supposed to enhance the resilience and trustworthiness of AI are essential aspects for the end-users. These aspects can help to develop more secure, explainable, and trustworthy AI-based systems and security solutions used by the end users. Besides, the proposed resilient accountable metrics can help to verify and validate the AI-based applications, which enhances the acceptances of users in these products. Furthermore, improved transparency and explainability will increase users' trust in AI-based systems and networks. Finally, users also insist that their privacy will be preserved so that none of the users' personal data (e.g., health records) might be aggregated, leaked, or sold against their will. Here, the privacy-preserving methods developed within SPATIAL can help to satisfy this inherent user requirement.

In conclusion, we state that the level of interest of End Users in the SPATIAL activities and outcomes is medium since most of the envisioned frameworks and metrics will be developed for developers of AI-based systems. In addition, their level of power needs to be considered high.

Level of Interest: medium

Level of Power: high

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

3.3.3 AUDITORS

Another group of stakeholders is represented by **Auditors.** As the name indicates, these are experts that audit the legal compliance as well as the correctness and security of technical systems and networks. Auditors have a deep understanding of both the legal aspects that need to be considered as well as technological aspects and standards applicable to applications and systems. As stated in Section 2, AI applications and systems are considered as opaque black-boxes, for which the concrete behaviour is untransparent and uncertain. This renders the verification and validation of AI-based systems difficult for auditors. Therefore, the accountable and explainable methods to be developed in SPATIAL can help the auditor to peek into the black-box and better understand the behaviour of AI-based systems. Also, the resilience and accountability metrics to be developed in SPATIAL can help to create the audit of AI-based systems and networks. Furthermore, the envisaged effective and practical adoption and adaptation guidelines to ensure streamlined implementation of trustworthy AI solutions, as well as the planned standardization activities regarding resilient and accountable AI methods are of particular interest to auditors. As a result, we consider the level of interest for auditors to be high. The SPATIAL outcomes may help auditors understand how AI-based systems work and conclude whether the systems' behaviour is compliant with the legal restrictions. This can also be a relevant aspect in the certification process of AI-based systems and applications. In the future, SPATIAL might provide fundamental tools helping auditors in their daily practice in understanding which functionalities concern legal aspects, which requirements are met, and how applications are secured. Therefore, we also consider the level of power of auditors as high.

Level of Interest: high

Level of Power: high

3.3.4 TESTERS

In addition to developers, **Testers** are also a relevant stakeholder group in SPATIAL. In general, testers examine the whole range of functionality of systems and applications, test every aspect of them thoroughly, assess their security attributes, and analyse their robustness and resilience. Especially for AI-based systems, testers face the same problems as auditors regarding the black-box behaviour of these systems. Testers need to understand the technical details of the AI-systems' implementation to identify errors and look for security issues (e.g., backdoors, possible privacy leaks, or other security threats). However, this is difficult for AI-based systems due to AI models' opaque decision-making process. Therefore, the explainability and accountability methods to be developed in SPATIAL can again help to understand the black-box behaviour of AI models and to understand AI systems better. As a result, these systems can be tested more precisely and subsequently improved in a more targeted manner. A further relevant aspect for testers is related to the resilience and accountability metrics to be developed in SPATIAL. In order to reach dependable measurements, testers need clear metrics and methods. Since SPATIAL will provide these, this leads to a high level of interest for testers. In conclusion, we classify the interest of developers in the SPATIAL activities and outcomes as high. However, since testers are a specialized target group, we consider their level of power to be medium.

Level of Interest: high

Level of Power: medium

3.3.5 SYSTEM OPERATORS

The **System Operator's** task is to implement, maintain, and improve AI-based systems and networks. Typically, they create and maintain complex IT systems by orchestrating many diverse components, of which only some are based on AI. Nevertheless, system operators are also a relevant stakeholder group in SPATIAL. They are interested in the security, robustness, and resilience of their complex systems and networks. In addition, legal aspects regarding the operation of their systems, such as the accountability of individual components or the compliance with data protection regulations (e.g., GDPR), also play a decisive role for system operators. Especially in connection with AI components, legal accountability and trustworthiness are a significant challenge for system operators since the decision-making process as well as the behaviour of individual AI components is often not comprehensible and uncertain. This hinders the standardization and certification process of their systems. Regarding these aspects, the privacy-preserving methods, resilience and accountability metrics, frameworks to increase security and explainability, and guidelines regarding trustworthy and secure AI investigated in SPATIAL can be of great importance for system operators. However, we classify the level of interest and the level of power of system operators just as medium. Although some components of their orchestrated complex IT systems are AI-based and thus the SPATIAL outcomes may be relevant for them, we still consider the developers, testers, and auditors in SPATIAL to be of greater importance. These stakeholders develop, test, and verify the AI-based components, so the outcomes of the SPATIAL project will be more impactful for these groups.

Level of Interest: medium

Level of Power: medium

3.4 REQUIREMENTS

In the subsequent section, we introduce general aspects and design principles to be considered when integrating and utilizing AI algorithms and frameworks for addressing security risks and threats to system architectures in the form of precise requirements assigned to several categories. In the following, individual subsections are dedicated to each of these categories, in which we discuss the relevance of the captured requirements and summarize them in the form of tables. In this context, it is important to note that in addition to the general aspects and design principles, we will also consider requirements specific to the four SPATIAL use cases.

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

Furthermore, we want to emphasize that although some requirements might look similar in character, they are formulated from different perspectives and with respect to different components.

3.4.1 SOFTWARE AND HARDWARE REQUIREMENTS

3.4.1.1 Software Requirements

When it comes to the federation and secure execution of AI algorithms, the SPATIAL project aims to leverage the edge-cloud continuum. For this reason, we will cover software requirements that include both edge and the cloud capabilities in the following. In this context, we assume that the cloud will be organized around a microservice-based architecture, while corresponding Edge and Trusted Execution Environment (TEE) agents will run locally and communicate with cloud components for synchronization, management, and authorization. In this scenario, we identified three main software categories of interest: cloud for microservices and federated AI flows orchestration, edge for data acquisition and edge management as well as software agents for running in the TEE to monitor the execution and provide the end-user access to results. The identified requirements are summarized Table 2 (see Appendix A).

Federated AI orchestration: Any microservices developed in the context of the SPATIAL project SHOULD be written as cloud native apps with Kubernetes as a default deployment orchestrator. Furthermore, the utilized Kubernetes orchestrator SHOULD have capabilities to enable federated AI execution similar to modern projects like Kubeflow [129].

Secure and confidential cloud execution: To enable confidential computing for AI-based systems deployed at the Edge or in the Cloud, software enablers for TEE SHOULD be used. Examples are Linux Foundation Confidential Computing Consortium projects like Enarx [130] or Veracruz [131]. These include libraries and agents that create Hardware Abstraction Layer (HAL) abstractions within the secure enclaves. Additionally, management of remote attestations are needed. Execution in secure enclaves will be managed by the software agents that will be running in enclaves. This agent will provide an interface for communication between end-user or edge and secure enclaves on the one side, and between enclaves and the cloud the other side. Since data is decrypted in the TEE, decryption keys must be sent using agent directly in the enclave. Additionally, data privacy might be in danger if the algorithm is not trustworthy and tries to leak the data from the enclave. Protection from these cases can be algorithm certification, differential privacy, or purely cryptographic methods like Fully Homomorphic Encryption.

Edge agent: Edge nodes SHOULD utilize a so-called Software Management Agent (SMA) that is deployed on edge nodes and is capable to exchange messages with the cloud control software for various purposes - IoT telemetry, edge orchestration, edge service monitoring, and so on. Monitoring of the processes on the edge should be reported via a defined edge-cloud telemetry

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

protocol. SMA should be capable of monitoring other processes and assessing their status. Moreover, it SHOULD be capable of receiving commands from the cloud and executing them in order to control other processes (e.g. AI algorithms running on the edge node). The edge agent is conceptually similar to the TEE agent, but instead of providing access to the execution from within the enclave, it provides an interface to control edge nodes as well as enables edge nodes to do data acquisition or, even, do a lightweight pre-processing to filter which data will go to the cloud. Examples are Mainflux Agent [132] and EdgeX Foundry SMA [133].

3.4.1.2 Hardware Requirements

Our exploration of SPATIAL considers a distributed environment that consist of heterogenous networks. We list the key features of the environment that we assume in our use cases. The results are finally summarized in Table 3 (see Appendix A) in the form of concrete requirements.

Distributed edge infrastructure: In order to support (hierarchical) federated learning applications, our SPATIAL research requires an edge infrastructure which involves user devices, edge, and core network where the AI/ML modules can be flexibly deployed.

Hardware acceleration on the Edge (GPU, FPGA, etc.): Traditional general-purpose Central Processing Units (CPUs) have limited processing power, and this is one of the biggest constrains to perform computationally expensive functions at the edge. Hardware acceleration devices help to improve the computational power. At one end there are Application Specific Integrated Chips (ASICs). These are processing units designed to perform applications specific tasks. Thus, they are highly energy efficient [93], but very expensive to produce as any change in the application needs production of new hardware. On the other end, we have programmable processing units such as Field Programmable Gate Arrays (FPGA). These provide flexibility as they have gate arrays which can be programmed to generate the logical hardware units. Thus, they are also easier to deploy than the ASICs. Another type of accelerators is the Graphics Processing Unit (GPUs), which can accelerate the processing of graphical data. These are also available for embedded devices, e.g., Nvidia Jetson Nano [94].

4G/5G networks: Our research on AI-based network security solutions assumes 4G/5G networks which consist of a radio access network and evolved packet cores or 5G core. The network can have user equipment that access IP services such as browsers, web applications, VoIP calls, etc.

IoT networks: The same line of work also covers IoT networks, which include a border gateway and a set of sensors that periodically transmit data, and a server that collects the data.

Constrained Nodes (e.g., TinyML): IoT nodes can be very constrained sensors. Nevertheless, because of their expected number and deployment in the field (physical systems), they are good candidates for the application of certain ML algorithms. A TinyML provides an example [92].

VoIP enabled end devices: Use Case 3 includes an intelligent emergency call management. It assumes that the emergency call can be initiated from various devices (e.g., smartphones, smart TVs, smartwatches, or desktop computers) via SIP and WebRTC.

Health sensors: For emergency detection, Use Case 3 also requires health sensors (e.g., smart T-shirts) that send their measurement data via the Bluetooth protocol to the smartphone.

Location Information Server: In the context of Use Case 3, the LIS ensures that the respective caller device can resolve the location of the emergency caller using either DHCP, LLDP-MED, or HELD protocol.

3.4.2 DATA REQUIREMENTS

The quality of AI applications relies heavily on the quality of the underlaying data. Hence, data and its processing are key elements in the training of robust and reliable AI models [95] [96]. Large amounts of data are typically required to train models with high levels of accuracy. Besides this, quality requirements of the data also play a significant role to improve the overall model performance. As a result, different data trade-offs have to be considered when training AI models. In this section, we highlight data requirements that need to be considered for improving the quality, reliability, security, and fairness of AI models. These requirements are also summarized in Table 4 in Appendix A. We focus on establishing general high-level requirements that can be applied not just to the use cases in the project, but also can be generalized across other applications that rely on AI.

Supply of high-quality fair and unbiased data: When it comes to training an accurate AI model, one of the most essential and critically important pre-requisite requirements for the AI practitioner is to have access to large volume of high-quality input data. Acquiring such data is a challenging task, since the term “high-quality” entails multiple sub-properties to which the data SHOULD adhere. Some of these properties include data representativeness, fairness, and bias. First, the training, validation, and testing data MUST be a good representative of the use case for which the model will be applied in practice. More specifically, the features and the labels MUST follow a similar distribution as the data encountered in the production environment and there SHOULD be a sufficient number of the edge and corner cases that could potentially occur in practice. Additionally, the data set MUST be fair and unbiased. This means that discrimination in the model predictions based on sensitive inputs (e.g. ethnicity, health, gender, religion, race etc.) MUST be prevented at all costs. As proposed by S. Hajian et al. [127], reducing the bias and increasing the data fairness can be achieved by utilizing a variety of discrimination discovery techniques and by performing fairness-aware data mining.

Consistent data preparation and management: In general, raw data is pre-processed to fit AI models. By defining standard practices for data search, ordering, and manipulation, it is possible then to guarantee that its input format structure and training process can be replicated easily. Data governance standards MUST be adopted at organization level, such that it is possible to

SPATIAL project is funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 101021808.

analyse data by different organizations, in order to ensure transparency [97]. In addition, storing practices should be also guaranteed to keep the consistency of raw data. FAIR data standard is a good example for raw data management [100].

Aggregation/Combination of different heterogeneous data sources: A single set of data is typically not enough to bootstrap an AI model [98] [99]. As a result, a combination of heterogeneous data from multiple sources MUST be adopted. Given the often-required large amount of data, the process of data combination cannot manual, but rather an automatic one, in which missing data, duplicated data conflicts and data inconsistencies are resolved at once.

Sensitive and private information must be extracted before training the model: Data removal is important to protect information of individuals (if any) when training the model [99]. Unfortunately, removing data implies increasing the sparsity of data, making it difficult for extracting learning representative features.

Processing pipelines for feature extraction and definition of input data formats: AI models do not learn from raw data. Instead, features and variables are identified and extracted from the raw data, such that input data formats can be defined. Different models can rely on different data formats. However, features can be established from analysing generic patterns in the data using processing pipelines that look for correlations, outliers, and transformations [95] [96] [97] [98][98] [99].

Bias analysis of data semantics: Since heterogeneous data is merged and transformed to produce big datasets to train AI models, a logical next step is to analyse that the semantics of the data has not changed, even after transformations [97]. Hence, after data is prepared, the pre-processed data SHOULD be analysed to ensure that data semantics are preserved. This is to guarantee that noise is not introduced in the data to produce bias and erroneous decisions.

Quantifying data quality for AI training: Performance of machine learning models is an indirect measure that determines the quality of the data [97]. Bias, variance, and irreducible errors are factors to quantify the quality of the AI models in decision making. While it is not possible to have a balanced trade-off between those factors, it is possible to tune an AI model to perform the best by linking performance to the analysis of those trade-offs. As a result, data quality SHOULD be measured by quantifying the performance of the AI model.

Enriching data quality: Multiple methods are available to improve the quality of data. Data properties such as sparsity and temporal locality can be easily overcome by relying on imputation and reconstruction methods [96] [98], e.g. compressive sensing, sparse coding, interpolation. Data MAY be enriched to allow AI models to provide optimal decision making.

Fixed data dimension for AI models: Besides relying on model performance metrics to indirectly measure data quality, a set of dimensions can be used instead to quantify data quality, e.g. accuracy, currency, and consistency. Available dimensions to model data quality can vary as there are different available definitions. Enforcing data standards to analyse quality of models is required [97]. Therefore, we recommend that data quality for AI training SHOULD be also explicitly defined by data dimensions.

Linking input data with prediction outputs: As mentioned, indirect performance metrics of models can be used to assess the input data. However, explicit linking between data input and output predictions of models SHOULD be considered for deriving quantifiable explanations to users.

Continuous data preparation and consistency: Modern federated learning methods facilitate continuous learning by aggregating data over time to models. Data consistency and integrity MUST be preserved [96] [99], meaning that data properties are not drastically modified/drift, e.g, variance, prediction accuracy and so on.

3.4.2.1 Privacy

The privacy of data is generally known to be the assurance that individuals get the control or influence of what details related to them may be collected and stored, and by whom and to whom the information may be disclosed [101]. Therefore, it is the capability that a person gets to seclude the information about themselves selectively. This is applicable for data privacy in AI as well, since data used for AI applications require privacy consideration to ensure privacy for data owners. The following summarises some identified requirements to be considered when maintaining the data privacy of AI systems and applications. The gathered requirements are also represented in Table 5 in Appendix A.

Identify the data category for privacy protection: According to GPDR definition Art. 4 [102], there are two categories of data considering the linkage with the data source: personal and non-personal data. The privacy requirements may vary based on the nature of the data. Therefore, one SHOULD consider the privacy considerations of data types that are input into AI models.

Privacy protection for Big Data: The privacy SHOULD be considered in each stage; data generation, data storage, and data processing [103]. For AI models, the prerequisites may be the privacy guarantees for each of these stages since the model's capabilities would depend significantly on the quality of collected data. If adversaries could poison the data in any of the mentioned states, it would be impacting the overall model's metrics such as accuracy, precision, and properties such as bias.

Privacy preserved data portability: In order to maintain the flexibility of data transferring without re-entering or unnecessary duplication of data for AI applications, it would be required to retain the portability of data to a certain extent[10]. However, during such requirements, the

[10] In this scenario, "data portability" refers to the concept described in the GDPR.

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

privacy of data may be properly evaluated. Furthermore, consent from data owners may be required during such instances.

Erasure and rectification of personal data: A person MUST be able to request to rectify incorrect or incomplete personal data or to erase them from the digital records [104]. This would be required for data privacy since the individuals who contribute to creating the dataset for AI models may need their data removed and not included in the model development process.

Assessing privacy attacks on AI system data: There are numerous types of attacks on AI models, which may reveal the privacy of subjects used to train the models or use the trained model [105]. The model may be subject to membership inference and attribute inference attacks, making data owners vulnerable, even when the original data is secured. Therefore, the possibility of privacy-related attacks on AI, system, and data MUST be assessed and protection or mitigation processes MUST be made.

Introducing privacy metrics: Identifying proper metrics would be necessary for AI privacy as it simplifies which actions to be taken to ensure the privacy level. It would help in defining which steps should be taken on privacy violations. Hence, we conclude that there SHOULD be proper metrics defined for privacy to support privacy protection measures

Privacy by design: By default, AI-based systems should safeguard data privacy requirements, even if end users pay less attention to privacy rights and threats. It is required to protect beforehand, not after a privacy breach [106].

Privacy versus performance: The data privacy-preserving techniques may cause performance degradation with the implementation in the real-world [107]. In this case, it could support evaluating the need for privacy and the costs involved when operating versus in a breach. As a result, trade-offs between model performance and privacy MAY be considered when implementing privacy.

3.4.3 MODEL REQUIREMENTS

The quality of ML models is conditioned by many requirements. First, it is highly dependent on data requirements such as the quantity and the quality of the data the model is trained with -i.e., requirements that were discussed in the previous section (see Section 3.4.2). Beyond data requirements, there are several model requirements that must be met during the design, implementation, and deployment of ML models. Some studies [108] and regulations [109] have already established initial lists of generic requirements for ML models. These are completed with methodologies for ML requirement engineering to defined use-case specific model requirements [110]. We summarize here the main ML model requirement that can be expected for Mobile Edge Systems, Cybersecurity Applications and Analytics, IoT, and eHealth. The results are also listed in Table 6 (see Appendix A).

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

Accuracy: Accuracy is a primary requirement that establishes if an ML model is usable to fulfil its primary purpose: it evaluates the correctness of its predictions. Accuracy is typically the first targeted requirement, which is algorithmically optimized during the training of the model. Therefore, we conclude that ML models MUST have a high accuracy. Precisely, the accuracy of models is assessed by comparing its predictions to the real labels of ground truth data. The performance of an ML model can be evaluated using several accuracy metrics such as precision, recall, Area Under the Curve (AUC), F-score, etc.

Generalizability: Generalizability is the second paramount requirement establishing that ML models MUST keep a high performance on unseen and unknown data. Generalizability ensures the reliability of predictions after the ML model is trained and deployed. Generalizability is often assessed by computing the accuracy of the ML model on new ground truth data that was not used during training. Several training and optimization strategies (e.g., regularization) can be used to enforce this requirement by design.

Maintainability: ML models need to preserve their good performance overtime. Monitoring of model performance should be possible and means to react to a performance decrease, e.g., model retraining, should be possible to execute. Maintainability establishes that ML models MUST be able to adapt to changes in their environment for preserving their good performance. It should be possible and easy to update and to retrain them to that end.

Response time / latency: ML models SHOULD be responsive and able to render their decision in a limited amount of time. This requirement is generic to many systems rather than specific to ML models. However, this requirement is critical for ML models used in autonomous systems, which require low latency and a short response time to take, e.g., safety critical decisions.

Fairness: Fairness is a recent requirement establishing that ML models do not reproduce human bias based on certain discriminatory features. This requirement is particularly important when data and features characterizing people are used and when ML models are used in decision making having a societal impact. Fairness ensures that characteristics deemed discriminatory like age, gender or race are not considered, even indirectly, by the ML model to render its decisions. As a result, we state that ML models SHOULD NOT output biased decisions.

Explainability: It is often unclear how results are output by ML models due to the oblivious decision process that many ML models implement. Explainability establishes that each prediction output by the model can be explained in a human intelligible manner that can rationalize its decision. This requirement is meant to increase the trust in the decision of ML models by making them understandable to people. Explainability can also serve to prove that the fairness requirement is met. Hence, ML models' predictions SHOULD be understandable by humans.

Transparency: The transparency requirement enforces that information about how ML models are built (training process, algorithm, training data used, features extracted, etc.) and the way they are used, are documented and available for auditing by relevant parties. It has a similar purpose as the explainability requirement: increasing trust in ML-based systems and providing proofs that requirements such as privacy and fairness are met. It also has the additional purpose of proving a legitimate use of the ML model. Therefore, the training, deployment, and usage of ML models MUST be documented and accessible.

Testability: ML models SHOULD be testable to verify that the decisions they output are compliant with the specification established during their design, much like software testing. Testability also covers vulnerability testing to assess the vulnerabilities of the ML model to adversarial ML attacks. The testability requirement is necessary to increase the confidence that the decision of the model will be reliable and its performance good when deployed.

Scalability and training performance: ML models SHOULD be trained (e.g. convergence time) in a reasonable amount of time and be scalable with respect to the amount of data. The scalability and training performance should also consider the scenarios of distributed model developments and models of multiple applications/services.

3.4.4 LEGISLATIVE REQUIREMENTS

The legislative requirements described below reflect both AI development and AI implementation. The first set of requirements is derived from the General Data Protection Regulation, as this is the most comprehensive regulatory framework currently in place. These are then followed by an overview of requirements that are likely applicable to SPATIAL outcomes when the Artificial Intelligence Act comes into action, which is projected for mid-2022. Table 7 in Appendix A summarizes the identified legislative requirements.

General Data Protection Regulation: The GDPR applies to the collection and processing of personal data concerning European citizens. Implemented in 2018, the General Data Protection Regulation is the most comprehensive privacy regulatory framework to date in the European Union (EU). The GDPR was designed to regulate how natural persons' data is to be handled by entities collecting and processing these personal data. It operates under seven principles: lawfulness, fairness, and transparency; accuracy, data minimisation, integrity, and confidentiality; storage limitation; purpose limitation; and accountability. For SPATIAL we are ensuring that every action taken is GDPR compliant. Here are discussed those actions.

Data subjects, those whose information is collected, must be informed about how the data is being collected, processed, and disseminated in a clear, accessible, easy-to-read, and understandable way. Furthermore, data subjects have the right to access their personal data within a month from the date they requested it, or to object to the processing of their data or any data that may concern them. Together with the information about the data subjects, SPATIAL shall provide details regarding the reason why that data is being processed, how the information is being used, and how long it will be stored.

Additionally, data subjects have the right to request their personal data to be erased in some circumstances. Some of the reasons may be in cases where data is no longer necessary or used for the purpose for which it was collected, if the data subject withdraws consent, if there is no legitimate interest from the legal person, or if it was unlawfully processed. SPATIAL must, in a timely manner, proceed with the deletion of these data.

When using data to perform automated decisions it is important that these are not solely based on an algorithmic decision, but also have a human component assessing the fairness of the output. Prior to collecting and processing personal information of European citizens, a data protection impact assessment (DPIA) and data management plan (DMP) should be performed. Both the DPIA and DMP are live documents and should be monitored and updated to reflect ongoing work and practices. Equally, all processes should be exhaustively documented, to ensure an auditable track record of every decision. By extension, this will ensure that data collection and processing begin and remain transparent and accountable. Such processes are already incorporated in the project; however, they should also be set up for the SPATIAL use cases.

The Artificial Intelligence Act: The European Commission is working towards the implementation of the Artificial Intelligence Act. The Act is projected to come into force in the second half of 2022. Its aim is to formalize the standards for trustworthy AI, including requirements for legal, ethical, and technical aspects for AI innovations. A key spearpoint is that AI should preserve democratic ideas, human rights, and the rule of law [111]. The Act sets out regulations that aim to minimize risks or potential negative consequences towards individuals or society due to the implementation of AI [112]. The Act takes a risk-based approach, meaning that the regulations, and how strict they are, differ based on the risk level of the AI solutions that are being developed. The four levels of risk are: (i) an unacceptable risk, (ii) a high risk, (iii) limited risk, and (iv) low or minimal risk [112] .

At the time of writing, the AI Act has yet to come into effect, but it is important to consider the potential implications for the project. Based on Annex III of the current AI Act text [113], which outlines what type of AI applications are high risk, we suggest considering this risk level for all SPATIAL work. This is due to the nature of SPATIAL work and potential outcomes, for example in relation to Use Case 3 “Accountable AI in Emergency eCall System.” This outcome has implications for emergency response, which is one of the high-risk criteria outlined in Annex III of the AI Act.

Adopting the framework for high-risk AI for all use cases will increase consistency, as well as a high level of transparency and accountability across the project. It will require the use case leads to create and maintain a risk management system; to test for risk identification and mitigation, including validation that the system will run consistently; to establish suitable data governance controls and ensuring that the used datasets for training, validation, and testing are complete, without errors, and representative. In addition, there should be comprehensive technical documentation that includes system architecture, algorithmic design, and specifications of the models. High-risk AI should be outfitted with automatic logging of events that complies with (EU) recognized standards. Finally, and related to requirements also set out in the GDPR, system output should be sufficiently transparent so that users can interpret the outcomes, and there should be human oversight over the AI at all times, which may include and override or off-switch capability.

Additional legislative requirements: Aside from these, the AI Act also strives for high levels of standardization across the EU, suggesting that project partners that lead the development of the AI revisit recognized standards on cybersecurity and/or sector-specific standards (telco, IT, and others) within the EU to ensure that their designs comply with these standards.

Finally, the AI Act has close connections to other legislation that may apply, such as the Data Governance Act, the Open Data Directive, the New Legislative Framework, and the new Data Act. The implications of such directives be explored further to ensure that SPATIAL remains in line with relevant legislation.

3.4.5 SECURITY REQUIREMENTS

In Section 2, we specified that AI-based systems are to be understood as a combination of traditional Information Technology (IT) components and specific AI-based components. Therefore, like any other computing system, also AI-based systems are exposed to security threats and cyber-attacks. However, for the traditional components, many research studies already exist that summarize existing security risks and threats and recommend appropriate countermeasures and design guidelines. Therefore, in order not to exceed the scope of this deliverable document, we will neglect security risks and threats specific to traditional components in this document. Instead, we limit ourselves to novel threats and risks for the AI-based components of such systems, for which the security of the underlying AI model is of utmost importance. In this context, many risks, threats, and adversarial attacks have been identified in recent years (see Section 2.3) aiming to attack or manipulate the AI models leveraged by AI-based systems. Hence, it is important for developers, testers, and operators of AI-based systems to be aware of these and consider them during the design, development, and deployment phase of AI-based systems. Therefore, we review potential security risks and threats for AI-based systems and identify corresponding requirements for dealing with them in the following. The resulting requirements are summarized in Table 8 in Appendix A.

Resilience against evasion attacks: One of the main attack vectors for AI models are the so-called *evasion attacks* (see Section 2.3). In this type of attacks, an attacker aims to create *adversarial examples* by carefully manipulating input data by adding small perturbation to it. The

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

goal of these adversarial examples is to cause the trained AI model to classify the manipulated data no longer correctly. As a consequence, this can result in serious security concerns since attackers could achieve to manipulate the behaviour of AI-based systems in a targeted manner. Therefore, AI-based systems MUST be resilient to this kind of attacks.

Resilience against data poisoning attacks: A further category of attacks against AI models is known as *data poisoning attacks* (see Section 2.3). Here, an attacker aims to manipulate the training data supplied at the training phase of an AI model. The goal of the attacker is to significantly decrease the performance (e.g. in terms of classification accuracy) of the trained AI models and to impair the normal functions of the AI system. Precisely, three concrete attacks are known that aim to contaminate the training data: *label flipping attacks*, *data injection attacks*, and *input manipulation attacks* (see Section 2.3). In order to mitigate these kinds of attacks, various *anti-poisoning techniques* such as *data sanitization* exist that can be applied to data supplied from untrusted sources. However, presenting such techniques fell out of scope of this work. In conclusion, we state that AI-based systems MUST be resilient to data poisoning attacks and MAY apply anti-poisoning methods to training data obtained from untrusted sources.

Resilience against backdoor attacks: Another type of existing attacks on AI models are *backdoor attacks*, in which AI models can be intentionally or unintentionally embedded with backdoors that could be triggered by the attacker. These backdoors are usually activated by the attacker through simple and often difficult to detect permutations of the input data that selectively and only under certain conditions change the behaviour of AI models. Since a backdoored model is designed to exhibit adversarial behaviour on inputs, which are only known to the attacker, this attack type is inherently hard to detect. Again, this poses serious risks in the operation of AI-based systems. Especially in the context of federated learning scenarios, backdoor attacks are of high relevance. In federated learning, clients provide model updates based on their private data to a central server that combines client-provided model updates to obtain a global model. By issuing carefully crafted updates, malicious clients may determine a backdoored global model, i.e., a model that assigns a wrong classification to all inputs with a certain (attacker-chosen) feature, while it behaves normally on inputs lacking such feature. For example, Bagdasaryan et al. [114] and Bhagoji et al. [115] demonstrate backdoor attacks by a single client in federated learning applications. Another backdoor attack on FL is the distributed backdoor attack (DBA) proposed by Xie et al. [116] in the context of backdoor-trigger attacks. DBA leverages multiple clients to submit poisoned updates containing a "trigger portion" each so that the resulting global model is sensitive to the combined trigger. To mitigate backdoor attacks and also poisoning attacks in the context of FL, several approaches already exist. The first proposal to mitigate poisoning attacks in FL is FoolsGold by Fung et al. [117], introducing the methodology of inspecting local updates and filtering out the suspicious ones. FoolsGold assumes that every class is represented in the data of some honest client, and it relies on the attacker operating through multiple clients. Li et al. [118] use spectral anomaly-detection methods to detect malicious updates in order to defeat both targeted and untargeted attacks. Nguyen et al. [119] present FLguard, a two-layer defence to filter out local updates with high backdoor impact and remove residual backdoors via clipping, smoothing, and noise addition. The private version of FLguard guarantees privacy but introduces considerable and costly changes to the FL process. Pillutla et al. [120] presents Robust Federated Aggregation (RFA), lifting the approach of robust distributed learning to the FL scenario. RFA seeks robustness against untargeted attacks degrading the overall classification accuracy. As a result of the discussion, we note that AI-based systems MUST be resilient against backdoor attacks.

Resilience against model extraction attacks: Another attack to be considered in the development and deployment of AI-based systems are the so-called *model extraction attacks* (see Section 2.3). For this kind of attack, an attacker observes the input, output, and other information, such as the parameters or training data of an AI model during the interference time. Subsequently, based on the gathered information, the attacker aims to reconstruct the AI model (e.g, by approximating a surrogate model) or craft adversarial examples that can be used to attack the model. Thus, besides security concerns, privacy issues as well as violations of intellectual properties can appear. Therefore, AI-based systems MUST be resilient to this kind of attack.

Resilience against data privacy attacks: Many AI-based systems employ AI services that use personal data (images, locations traces, web site access logs, etc.) at a large scale, for example, personalized recommendation services, targeted advertisement, credit assessment, hiring, to name a few. As such systems use highly privacy-sensitive data, potential attacks to the systems can lead to serious privacy violation. Therefore, data privacy attacks need to be considered in the design and development stage of AI-based systems and it must be guaranteed that AI-based are resilient against these. Precisely, we identified three variants of data privacy attacks to be considered: The *data reconstruction attack* (DRA) aims at reconstructing original input data based on the observed model or its gradients. For instance, based on aggregated statistics about a class of people, a DRA attack attempts to unfold the statistics and identify the raw data about an individual that was used to compute the statistics: for example, the age of individuals from an aggregated age statistic of a census. Another type of data privacy attacks are *property interference attacks* (PIAs). The goal of PIAs is to infer the value of private properties in the input data. Assuming an access to a trained model, an attack inspects the model and tries to figure out a global statistic about the training set: for example, an average score of the training set is under a particular threshold value. Lastly, *membership interference attacks* (MIAs) must be considered. The purpose of MIAs is to learn whether specific data instances are present in the training dataset. Again, assuming an access to a trained model, an attack could identify whether an individual belongs to a specific class or group. Since all of these three presented attacks can result in serious privacy violations, AI-based systems MUST be resilient against them and the ML model SHOULD NOT leak information about its training data. Technical solutions like privacy-preserving training and differential privacy are typically used to meet this requirement.

Resilience against attacks in online learning scenarios: AI based systems that apply online learning strategies can become targets of poisoning attacks applied during the operational

SPATIAL project is funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 101021808.

phase. In this scenario, the system under attack continues learning during the deployment phase by iteratively updating and retraining the deployed AI model with the data gathered during the operational phase. An attacker can exploit this behaviour by simply using the deployed model and supplying it with poisoned data. Hence, the attacker might be able to slowly retrain the attacked model in a targeted manner. Therefore, AI-based systems that apply online learning strategies SHOULD be resilient against such kind of attacks. Again, anti-poisoning techniques can be applied to mitigate these attacks.

Resilience against (D)DoS attacks: Furthermore, an attacker may aim to perform denial of service attacks on AI-based components by sending it a significant high amount or fairly complex requests to handle. This may result in an overload of the corresponding component, which in contrast results to a denial of service, in which the attacked component is unusable for its intended purpose. For AI-based components integrated into distributed (network) environments, the same risk applies also to distributed denial of service attacks (see Section 2.2.2). Hence, it should be guaranteed that AI-based systems are resilient against DoS and DDoS attacks.

Data protection in the operational phase: In addition to the privacy attacks mentioned above, it must also be ensured that AI-based systems do not intentionally or unintentionally leak private or confidential data during the operational phase (e.g. in the form of logging or auditing outputs) that can be used by an attacker. This is the only way to ensure holistic and end-to-end data privacy in such systems. Therefore, AI-based systems should be extensively screened for unintended information leaks prior to deployment. In conclusion, we state that AI-based systems that deal with sensitive or confidential data MUST preserve the confidentiality of the data during the operational phase.

3.4.6 USABILITY REQUIREMENTS

In the area of artificial intelligence and machine learning, AI-based systems exhibit different usability challenges and requirements than traditional human-computer interaction systems, making it difficult to rely on established usability guidelines and best practices. For example, the unpredictable, non-deterministic, and opaque black-box behaviour of AI-based systems can confuse users, decreasing their acceptance of and trust in these systems [121]. Although aware of these challenges, most of the research and development efforts in recent years focussed on improving the performance of AI models, e.g. in terms of classification accuracy. Increasing and fine-tuning the usability of AI-based applications and systems was often neglected [122]. Therefore, these usability concerns will be addressed from the beginning of the SPATIAL project and taken into account in the realization of the project objectives. Accordingly, we will discuss identified usability requirements that will be important in the context of SPATIAL in the following. These requirements are also summarized in Table 9 (see Appendix A). In this context, it should be mentioned that we will often present high-level requirements since the concrete level of usability challenges and requirements depends on the degree of automation, technical and domain expertise of users, and the associated risk of concrete AI-based systems [123].

Easy-to-use interfaces: Like every other system, AI-based systems MUST provide comprehensible, uniform, and easy-to-use interfaces (e.g., graphical user interfaces and API interfaces) for their users. By providing easy-to-use interfaces, users will accept and use AI-based systems in the long term.

Documentation and Help: Detailed and comprehensive documentation is of great importance for AI-based systems due to their probabilistic and opaque behaviour. Therefore, an AI-based system MUST be documented in detail. This documentation should include extensive information about the data used, the training process, as well as information about the utilized AI models and their deployment. The technical and domain expertise of the users should also be considered. Such documentation could increase users' understanding, acceptance, and trust in the often complex and untransparent AI-based systems. In addition, an AI-based system SHOULD have functionalities that guide users and provide help in case of problems.

Consistency and specification of decisions, outputs, and interfaces: Due to their non-deterministic and probabilistic behaviour, AI-based systems can generate inconsistent and confusing outputs [121]. For example, an AI model can generate different outputs and decisions for the same input. Furthermore, serious usability concerns can arise if the format of the output of the AI model changes or the model outputs auxiliary information instead of predictions [123]. Both aspects can cause interpretation difficulties and confuse users, which can lead to reduced acceptance of and trust in the system. Therefore, all decisions and outputs of AI-based systems SHOULD be as consistent as possible and follow pre-specified and interpretable formats. In addition, all user interfaces SHOULD be consistent and unified in order to increase the usability of the system.

Provide explanations: As already pointed out several times in this document, AI-based systems often exhibit a non-deterministic and opaque black-box behaviour. Even domain experts have difficulties understanding and explaining individual AI decisions for such systems. However, understanding and being able to explain is a fundamental requirement to gain trust and acceptance in any AI-based system [122] [124]. Therefore, AI-based systems MUST provide explanations for individual decisions of the deployed AI models. These explanations have to be adapted to the respective technical expertise and domain knowledge of the users. Furthermore, the explanations should be of high-quality and adapted to the respective use case. Suitable XAI methods should be identified and integrated into the AI-based system (see Appendix B). Only when individual AI decisions are accompanied by understandable and helpful explanations, complex and opaque AI-based systems can find trust and acceptance of the users in the long term. In this context, we want to mention the work of Liao et al. [125], which provides a question-driven framework to identify the users' needs in the context of explainable AI.

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

Provide information about the effects of the decision: Due to the often non-deterministic and opaque behaviour of AI-based systems, these systems SHOULD provide information about the effect of concrete AI decisions to the users. As mentioned above, many users lack an understanding of how AI-based systems work, which means that individual decisions cannot be comprehended. To support users in such a situation, remove the fear of erroneous and harmful consequences, and give them the feeling of control over the situation, such information about the effects of the decision is helpful. This feature could also increase the users' trust in the AI-based systems and their decisions.

Provide help in case of errors: Besides the explanations of the decision-making and the corresponding information about the effects, an AI-based system MAY also provide functionalities that offer support to the user in case of system errors. This could reduce the users' fear of possible erroneous behaviour of the probabilistic AI models and increase the acceptance and trust in the system.

Enable the correction of AI decisions: AI-based systems are often used to support human domain experts in their decision-making process. Therefore, Amershi et al. [121] suggest that users of AI-based systems SHOULD also be able to identify, report, and correct mistakes in the decision-making of AI models. Again, such a mechanism increases the trust of domain experts in the system while helping to improve the performance and reliability of the system.

Notify users about changes: As often mentioned in the previous sections, AI-based systems need to be dynamically adapted and improved. Since this may significantly change the behaviour of AI models and thus the entire system, users of AI-based systems MUST be informed about every system update. Afterwards, they can familiarize themselves with the system's adaptations and possible behavioural changes.

Update and adapt cautiously: Among their 18 proposed generally applicable design guidelines for human-AI interaction, Amershi et al. [121] also mention the importance of updating and adapting AI-based systems carefully and successively. Since, as described in Section 2, AI-based systems lack sophisticated specification, verification, and testing capabilities, only minor and incremental system changes SHOULD be applied. This should avoid disruptive and erroneous modifications and ensure the continuous usability of the system.

3.4.7 ACCESSIBILITY REQUIREMENTS

Accessibility is a key requirement in ensuring transparency and explainability beyond functional requirements. This is different from usability since it is not a requirement for the use of the algorithm, but it relates to the trustworthiness of AI solutions to the end-user and the broader lay community. This should be established through a ‘contract’ [126] embedded in the documentation that clarifies in accessible language and/or visuals the following aspects:

SPATIAL project is funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 101021808.

- **Purpose**: What is the purpose of the AI solution. This is an essential requirement to prevent function creep and misuse.
- **Risks:** What potential risks and impacts are foreseeable from the AI solution, including bias and discrimination. This is a relevant requirement for a transparent and informed use.
- **Process:** What are the key aspects of the AI solution's functioning. A clear explanation of the process means that the outputs of the solution will also be easier to interpret in context.
- **Accountability**: Who is accountable for the various components of the AI solution. While AI is often attributed to agent entities, a clear accountable entity is a necessary part of responsible deployment.

These aspects should be provided in a form that is *accessible without extensive background knowledge in ML and AI* (i.e. clear to lay persons). Furthermore, they should be *mindful of physical, mental, social and cultural vulnerabilities of users* (i.e. accessible to a wide and diverse range of individuals). The requirements discussed in this section are listed in Table 10 (see Appendix A).

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

4 OUTLOOK: DERIVING DESIGN GOALS FOR THE SPATIAL EXPLANATORY AI PLATFORM

One of the main objectives of the SPATIAL project is the design and development of an Explanatory AI platform, a tool for enhancing and quantifying the quality and trustworthiness of AI-based systems. The Explanatory AI platform will be designed to enable stakeholders to understand how an AI system makes predictions by offering methods that provide clear and understandable explanations of and the reasoning behind the decision-making of AI models. Furthermore, the platform also aims to translate the quality of AI systems into quantifiable data derived from accountability metrics as well as data quality estimates identified in the context of SPATIAL. Thus, the platform will enable stakeholders to enhance and evaluate the different properties of an AI system, such that it is possible to account its overall execution pipeline. In this chapter, we will explore how the requirements identified in this document can be used to derive concrete design goals and the expected functionality of the Explanatory AI platform. In this context, we want to emphasise that the aspects discussed in this chapter should not be understood as an approach for designing and modelling the SPATIAL platform. Instead, this chapter discusses how the presented requirements can be used as a starting point for developing the Explanatory AI platform. The technical definition of the SPATIAL platform including the design and specification of a particular distributed AI architecture with its components, properties, processes, and flows is carried out as part of the technical activities in other work packages.

As an illustrative example, the discussions and design guidelines provided in this document reveal that explainability, transparency, and accountability are crucial aspects of AI-based systems that can significantly impact their adoption and impact on society. These aspects are essential for building trust, understanding, and acceptance among users, especially in high-stake domains such as Mobile Edge Systems (e.g. 5G Services), Cybersecurity Applications and Analytics, IoT, and eHealth. The requirements and design guidelines gathered in this document clearly reflect these findings. Table 1 summarizes a collection of requirements that indicate the need for transparency and explainability for AI-based systems and, thus, are highly relevant to the SPATIAL use cases. Hence, the Explanatory AI platform should consider these formulated needs and offer appropriate tools and services to allow developers of AI-based systems to address these.

Therefore, the requirements listed in Table 1 indicate that the platform should be designed to provide insights into how an AI-based system arrives at its decisions. The platform should be designed to provide clear and understandable explanations of the decision-making of AI models by linking its inputs and outputs and thus enabling stakeholders to understand the reasoning behind the system's decisions. As discussed in this document, XAI methods can address these concerns by providing human-interpretable explanations for the decision-making of AI models. These XAI methods allow stakeholders to evaluate the decision-making process and identify any

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

biases, errors, or unfairness, which allows for preventing and addressing any potential ethical or legal issues in AI models. Therefore, the platform should offer services which employ XAI methods to generate explanations for individual decisions of AI models. Thereby, the explanations provided by these services must be generated by the most suitable XAI methods for an individual use case. As a consequence, the platform must offer multiple XAI methods. In this context, the usability requirements discussed in Section 3.4.6 impose additional demands on the explanations provided by the platform. As discussed in this section, explanations provided by the platform must adapt to the technical and domain expertise of the users, in order to achieve a high level of comprehensibility and interpretability. Therefore, the services deployed at the platform must provide explanations for individual decisions of AI models using comprehensible, uniform, and adaptive interfaces & explanations. Finally, the platform should also be designed to meet relevant data privacy and security regulations, such as the General Data Protection Regulation (GDPR).

The explainability requirements presented in Table 1 represent only a limited selection of aspects from which design goals for the platform can be derived. In this document, we identified other requirements the SPATIAL use cases should consider, and the platform should address. Precisely, all requirements allocated in Appendix A with the relevance *"Relevant for SPATIAL platform components"* should be reflected in the design of the platform. In this context, we want to mention that designing the explanatory AI platform involves a complex and iterative process that requires careful consideration of the problem being addressed, the data being used, and the interface provided to users. By considering the presented relevant requirements and following the discussed best practices in the field of AI, it is possible to create a platform that enhances transparency and explainability by providing users with clear and concise explanations of the model's decision-making process. To achieve this, the Explanatory AI platform should offer services that provide appropriate and adaptive explanations as well as actionable insights and quantifiable data that can be used to evaluate the quality of AI-based systems. This, in turn, can lead to increased accountability and transparency in the development and deployment of AI systems, which is crucial for building trust in AI and ensuring that it benefits society as a whole.

TABLE 1: SELECTION OF REQUIREMENTS RELEVANT FOR DERIVING DESIGN GOALS AND IDENTIFIYNG REQUIRED FUNCTIONALITY OF THE SPATIAL EXPLANATORY AI PLATFORM (SEE APPENDIX A)

Identifier	Software Requirements	Priority	Implemented By
DAT.RQ.13	Pre-processed input data SHOULD be linked with prediction outputs of AI models to derive quantifiable explanations to users.	SHOULD	Platform, UC2, UC3
MOD.RQ.6	ML models' predictions SHOULD provide high-level of explainability and should be understandable by humans.	SHOULD	Platform, UC2, UC3
LEG.RQ.10	According to the AI Act, AI MAY need to be designed with sufficient transparency to allow users to interpret the system's output.	MAY	Platform, UC2, UC3
USB.RQ.6	AI-based systems MUST provide explanations for individual decisions of the deployed AI models.	MUST	Platform, UC2, UC3
PRV.RQ.6	There SHOULD be proper metrics defined for privacy to support privacy protection measures	SHOULD	Platform
SW.RQ.1	Microservices developed in the context of the SPATIAL project SHOULD be written as cloud native apps with Kubernetes as a default deployment orchestrator.	SHOULD	Platform, UC1 UC2, UC3, UC4

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

5 CONCLUSIONS AND NEXT STEPS

Within this deliverable, we have pursued to goal to identify an initial set of requirements for the SPATIAL framework and belonging tools towards enabling the utilization of accountable and explainable AI/ML algorithms for the purpose of enhancing cybersecurity in modern system and network architectures. In order to achieve this goal, a set of important terms were defined - based on literature review - in order to establish the theoretical background of the project. These included terms such as *accountability*, *explainability*, *interpretability*, *resilience*, *transparency* and further, which are of paramount importance for the basic understanding of the project goals and emerging design principles and architectures.

Having reviewed some basic explainable AI methods as well as the general challenges to system architectures in key modern technological domains (such as 5G/6G, IoT, Edge Intelligence ...), we performed a deep dive into the reasoning for the extraction and cataloguing of tangible requirements of relevance for the SPATIAL project. This includes the detailed description of the four basic SPATIAL use cases, namely: (1) the utilization of privacy preserving AI in the cloud-fog-edge continuum, (2) improving the explainability, resilience and performance of cybersecurity in 4G/5G/6G and IoT networks, (3) the utilization of accountable AI in next generation emergency communication and (4) resilient cybersecurity analysis based on machine learning models. These four use cases and the belonging security and threat analysis are one of pillars for extracting and listing specific needs of modern system architectures in relation to the application of ML models for cybersecurity. Thereby, we discuss and define the specific stakeholders - such as end users, developers, testers, system operators and others - which are relevant for the emerging SPATIAL eco-system and bring in a special view on the overall framework and tools to emerge in the scope of the project. Finally, all these discussions are used as the foundation for cataloguing specific tangible requirements, which are classified as follows: software and hardware requirements, data requirements, model requirements, legislative requirements, security requirements, usability and finally accessibility requirements.

The above listed contributions provide an initial analysis and set the way forward for the SPATIAL project as a whole. The defined aspects can help to develop more secure, explainable, and trustworthy AI-based systems and security solutions. They aim at providing realistic guidelines for developers and operators on how to design, deploy, and modify AI-based systems, in order to provide streamlined application of secure, more transparent, explainable, and trustworthy AI.

The current catalogue constitutes just the first set of requirements within SPATIAL. More are expected to follow and will be continuously updated as the project with its algorithmic frameworks and tools progresses. Beyond the defined initial guidelines and aspects to consider, the project will use the current contributions to advance in the following directions: (1) definition of concrete resilience and explainability measures/metrics, (2) research and insights about the accountability and explainability of common AI/ML algorithms and design principles, (3) assessment of AI-based systems (reliability, limitations, etc) as well as (4) metrics for determining what attackers can accomplish and with what resources and capabilities. Another main challenge to be addressed by the SPATIAL consortium is (5) the technical definition of the SPATIAL framework/platform. This includes the design and specification of a particular distributed AI architecture with its components, properties, processes and flows that can operate in a modern networked environment and can easily be embedded in existing network and systems architectures, thereby ensuring the accountability and explainability of the applied ML models. By agilely incorporating the feedback and insights obtained from the just-mentioned technical activities as well as the realization of the SPATIAL use cases and platform, the requirements and design guidelines provided in this document will be refined and improved. An updated and final catalogue of requirements and design guidelines will be provided in deliverable D1.3 "*Final Requirements Analysis for AI towards Addressing Security Risks and Threats to System and Network Architectures*".

REFERENCES

[1] Bovens, M. (2007). Analysing and Assessing Accountability: A Conceptual Framework. European Law Journal.

[2] European-Commision. (2019). Ethics guidelines for trustworthy AI. https://doi.org/10.2759/346720

[3] Kacianka, S., & Pretschner, A. (2021). Designing Accountable Systems Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,

[4] Reddy, E., Cakici, B., & Ballestero, A. (2019). Beyond mystery: Putting algorithmic accountability in context. Big Data & Society, 6(1). https://doi.org/10.1177/2053951719826856

[5] Wieringa, M. (2020). What to account for when accounting for algorithms Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency

[6] W. Samek, T. Wiegand, en K.-R. Müller, "EXPLAINABLE ARTIFICIAL INTELLIGENCE: UNDERSTANDING, VISUALIZING AND INTERPRETING DEEP LEARNING MODELS", 2017.

[7] A. Adadi en M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)", IEEE access, vol 6, bll 52138–52160, 2018.

[8] A. Rosenfeld en A. Richardson, "Explainability in human--agent systems", Autonomous Agents and Multi-Agent Systems, vol 33, no 6, bll 673–705, 2019.

[9] NIST Definition of "resilience": https://csrc.nist.gov/glossary/term/resilience, as of date 29.03.2022

[10] Weller, Adrian. "Challenges for transparency." (2017).

[11] Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. "Transparent, explainable, and accountable AI for robotics." Science robotics 2, no. 6 (2017)

[12] Thelisson, Eva. "Towards Trust, Transparency and Liability in AI/AS systems." In IJCAI, pp. 5215-5216. 2017.

[13] Chromik, Michael, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek. "Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems." In IUI workshops, vol. 2327. 2019.

[14] Van Steen, Maarten, and Andrew S. Tanenbaum. Distributed systems. Leiden, The Netherlands: Maarten van Steen, 2017.

[15] Binder, Alexander, Wojciech Samek, Grégoire Montavon, Sebastian Bach, and Klaus-Robert Müller. "Analyzing and validating neural networks predictions." In Proceedings of the ICML 2016 Workshop on Visualization for Deep Learning. 2016.

[16] Vilone, Giulia, and Luca Longo. "Explainable artificial intelligence: a systematic review." arXiv preprint arXiv:2006.00093 (2020).

[17] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ACM Trans. Interact.

Intell. Syst. 11, 3–4, Article 24 (December 2021), 45 pages. DOI:https://doi.org/10.1145/3387166

[18] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135–1144. DOI: https://doi.org/10.1145/2939672.2939778

[19] NIST Definition of "trustworthiness": https://csrc.nist.gov/glossary/term/trustworthiness, as of date 29.03.2022

[20] European Commission, Directorate-General for Communications Networks, Content and Technology, Ethics guidelines for trustworthy AI, Publications Office, 2019, https://data.europa.eu/doi/10.2759/177365

[21] John Alexander, "Risk, Threat, or Vulnerability? How to Tell the Difference", online: https://www.kennasecurity.com/blog/risk-vs-threat-vs-vulnerability/, as of date 29.03.2022

[22] Y. Cherdantseva, P. Burnap, A. Blyth, P. Eden, K. Jones, H. Soulsby, K. Stoddart, "A review of cyber security risk assessment methods for SCADA systems", Comput. Secur., 56 (2016), pp. 1-27

[23] NIST Definition of "risk": https://csrc.nist.gov/glossary/term/risk, as of date 29.03.2022

[24] Stefan Fenz and Andreas Ekelhart. 2009. Formalizing information security knowledge. In Proceedings of the 4th International Symposium on Information, Computer, and Communications Security (ASIACCS '09).

[25] Data Privacy: https://gdpr.eu/data-privacy, as of date 29.03.2022

[26] Personal Data: https://gdpr-info.eu/issues/personal-data, as of date 29.03.2022

[27] M. Hachimi, G. Kaddoum, G. Gagnon, en P. Illy, "Multi-stage jamming attacks detection using deep learning combined with kernelized support vector machine in 5g cloud radio access networks", in 2020 International Symposium on Networks, Computers and Communications (ISNCC), 2020, bll 1–5.

[28] L. Gavrilovska, V. Rakovic, en D. Denkovski, "From Cloud RAN to Open RAN", Wirel. Pers. Commun., vol 113, no 3, bll 1523–1539, 2020.

[29] KrebsOnSecurity hit with record DDoS: https://krebsonsecurity.com/2016/09/krebsonsecurity-hit-with-record-ddos/, as of date 29.03.2022

[30] IoT Cyberattacks Escalate in 2021, According to Kaspersky, 2021 https://www.iotworldtoday.com/2021/09/17/iot-cyberattacks-escalate-in-2021-according-to-kaspersky/, as of date 29.03.2022

[31] Internet of Threats: IoT Botnets Drive Surge in Network Attacks: https://securityintelligence.com/posts/internet-of-threats-iot-botnets-network-attacks/, as of date 29.03.2022

[32] Log4Shell - What You Need To Know: https://www.randori.com/log4j/, as of date 29.03.2022

[33] E. de S. Nascimento, I. Ahmed, E. Oliveira, M. P. Palheta, I. Steinmacher, and T. Conte, 'Understanding Development Process of Machine Learning Systems: Challenges and Solutions', in 2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), Sep. 2019, pp. 1–6. doi: 10.1109/ESEM.2019.8870157.

[34] S. Martínez-Fernández et al., 'Software Engineering for AI-Based Systems: A Survey', ArXiv210501984 Cs, Sep. 2021, Accessed: Feb. 09, 2022. [Online]. Available: http://arxiv.org/abs/2105.01984

[35] H. Belani, M. Vukovic, and Ž. Car, 'Requirements Engineering Challenges in Building AI-Based Complex Systems', in 2019 IEEE 27th International Requirements Engineering Conference Workshops (REW), Sep. 2019, pp. 252–255. doi: 10.1109/REW.2019.00051.

[36] L. Fischer et al., 'AI System Engineering—Key Challenges and Lessons Learned', Mach. Learn. Knowl. Extr., vol. 3, no. 1, Art. no. 1, Mar. 2021, doi: 10.3390/make3010004.

[37] L. E. Lwakatare, A. Raj, I. Crnkovic, J. Bosch, and H. H. Olsson, 'Large-scale machine learning systems in real-world industrial settings: A review of challenges and solutions', Inf. Softw. Technol., vol. 127, p. 106368, Nov. 2020, doi: 10.1016/j.infsof.2020.106368.

[38] D. Silver et al., 'Mastering the game of Go with deep neural networks and tree search', Nature, vol. 529, no. 7587, pp. 484–489, Jan. 2016, doi: 10.1038/nature16961.

[39] M. Felderer and R. Ramler, 'Quality Assurance for AI-Based Systems: Overview and Challenges (Introduction to Interactive Session)', in Software Quality: Future Perspectives on Software Engineering Quality, vol. 404, D. Winkler, S. Biffl, D. Mendez, M. Wimmer, and J. Bergsmann, Eds. Cham: Springer International Publishing, 2021, pp. 33–42. doi: 10.1007/978-3-030-65854-0_3.

[40] J. Lin, L. Dang, M. Rahouti, and K. Xiong, 'ML Attack Models: Adversarial Attacks and Data Poisoning Attacks', ArXiv211202797 Cs, Dec. 2021, Accessed: Feb. 17, 2022. [Online]. Available: http://arxiv.org/abs/2112.02797

[41] A. Oseni, N. Moustafa, H. Janicke, P. Liu, Z. Tari, and A. Vasilakos, 'Security and Privacy for Artificial Intelligence: Opportunities and Challenges', ArXiv210204661 Cs, Feb. 2021, Accessed: Feb. 17, 2022. [Online]. Available: http://arxiv.org/abs/2102.04661

[42] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, 'Towards the Science of Security and Privacy in Machine Learning', ArXiv161103814 Cs, Nov. 2016, Accessed: Oct. 19, 2021. [Online]. Available: http://arxiv.org/abs/1611.03814

[43] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, 'SoK: Security and Privacy in Machine Learning', in 2018 IEEE European Symposium on Security and Privacy (EuroS P), Apr. 2018, pp. 399–414. doi: 10.1109/EuroSP.2018.00035.

[44] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He, 'A Survey of Human-in-the-loop for Machine Learning', ArXiv210800941 Cs, Nov. 2021, Accessed: Feb. 16, 2022. [Online]. Available: http://arxiv.org/abs/2108.00941

[45] Z. Zhang et al., 'Artificial intelligence in cyber security: research advances, challenges, and opportunities', Artif. Intell. Rev., Mar. 2021, doi: 10.1007/s10462-021-09976-0.

[46] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, 'Artificial Neural Networks-Based Machine Learning for Wireless Networks: A Tutorial', IEEE Commun. Surv. Tutor., vol. 21,

no. 4, pp. 3039–3071, 2019, doi: 10.1109/COMST.2019.2926625.

[47] V. Sze, Y.-H. Chen, J. Emer, A. Suleiman, and Z. Zhang, 'Hardware for machine learning: Challenges and opportunities', in 2017 IEEE Custom Integrated Circuits Conference (CICC), Apr. 2017, pp. 1–8. doi: 10.1109/CICC.2017.7993626.

[48] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI", Information Fusion, vol 58, bll 82–115, 2020.

[49] A. Singh, S. Sengupta, en V. Lakshminarayanan, "Explainable deep learning models in medical image analysis", Journal of Imaging, vol 6, no 6, bl 52, 2020.

[50] S. M. Lundberg und S.-I. Lee, „A unified approach to interpreting model predictions", in Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, Dezember 2017, S. 4768–4777.

[51] Sandra Wachter, Brent D. Mittelstadt, Chris Russell: Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. CoRR abs/1711.00399 (2017)

[52] C. Molnar, "Interpretable machine learning," Christoph Molnar, [Online]. Available: https://christophm.github.io/interpretable-ml-book/. [Accessed: 22-Mar-2022].

[53] Ramaravind Kommiya Mothilal, Amit Sharma, Chenhao Tan: Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. CoRR abs/1905.07697 (2019); https://github.com/interpretml/DiCE

[54] L. Breiman, "Random Forests", Machine Learning, 45(1), 5-32, 2001.

[55] Van der Maaten, L. and Hinton, G., 2008. Visualizing data using t-SNE. Journal of machine learning research, 9(11).

[56] Hinton, Geoffrey E., and Sam Roweis. "Stochastic neighbor embedding." Advances in neural information processing systems 15 (2002).

[57] Wattenberg, et al., "How to Use t-SNE Effectively", Distill, 2016. http://doi.org/10.23915/distill.00002

[58] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, Klaus-Robert Müller: Layer-Wise Relevance Propagation: An Overview. Explainable AI 2019: 193-209

[59] Montavon, G., Lapuschkin, S., Binder, A., Samek, W. and Müller, K.R., 2017. Explaining nonlinear classification decisions with deep taylor decomposition. Pattern recognition, 65, pp.211-222.

[60] Vilone, Giulia, and Luca Longo. "Explainable artificial intelligence: a systematic review." arXiv preprint arXiv:2006.00093 (2020).

[61] Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Springer, Cham.

[62] Bolei Zhou, A. K., Agata Lapedriza, Aude Oliva, Antonio Torralba "Learning Deep Features for Discriminative Localization."

[63] Li, X.-H., et al. (2020). "A Survey of Data-driven and Knowledge-aware eXplainable AI."

IEEE Transactions on Knowledge and Data Engineering: 1-1.

[64] Min Lin, Q. C., Shuicheng Yan (2014). "Network In Network."

[65] Ramprasaath R. Selvaraju, A. D., Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, Dhruv Batra (2017). "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization."

[66] Ramprasaath R. Selvaraju, A. D., Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, Dhruv Batra (2017). "Grad-CAM: Why did you say that?".

[67] Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." Annals of statistics (2001): 1189-1232.

[68] S. Bradner, “RFC 2119 - Key words for use in RFCs to Indicate Requirement Levels,” RFC Editor, RFC2119, Mar. 1997. doi: 10.17487/rfc2119.

[69] Robin C Geyer, Tassilo Klein, and Moin Nabi. 2017. Differentially private federated learning: A client level perspective. In NIPS Workshop: Machine Learning on the Phone and other Consumer Devices.

[70] OpenMined: https://www.openmined.org, as of date 29.03.2022

[71] Stacey Truex, Ling Liu, Ka-Ho Chow, Mehmet Emre Gursoy, and Wenqi Wei. 2020. LDP-Fed: Federated Learning with Local Differential Privacy. In 3rd International Workshop on Edge Systems, Analytics and Networking (EdgeSys).

[72] Lumin Liu, Jun Zhang, S.H. Song, and Khaled B. Letaief. 2020. Client-Edge-Cloud Hierarchical Federated Learning. In IEEE International Conference on Communications.

[73] Mehdi Salehi Heydar Abad, Emre Ozfatura, Deniz Gunduz, and Ozgur Ercetin. 2020. Hierarchical Federated Learning Across Heterogeneous Cellular Networks.

[74] Christopher Briggs, Zhong Fan, and Peter Andras. 2020. Federated learning with hierarchical clustering of local updates to improve training on non-IID data. https://arxiv.org/pdf/2004.11791.pdf. (2020).

[75] Stacey Truex, Nathalie Baracaldo, Ali Anwar, Thomas Steinke, Heiko Ludwig, Rui Zhang, and Yi Zhou. 2019. A hybrid approach to privacy-preserving federated learning. In Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security.

[76] Benjamin Zi Hao Zhao, Mohamed Ali Kaafar, and Nicolas Kourtellis. 2020. Not one but many Tradeoffs: Privacy Vs. Utility in Differentially Private Machine Learning. In ACM Cloud Computing Security Workshop (CCSW)

[77] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konecny, Stefano Mazzocchi, H. McMahan, Timon Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. 2019. Towards Federated Learning at Scale: System Design. In 2nd SysML Conference.

[78] Camps-Mur et al, “AI and ML – Enablers for Beyond 5G Networks.” 5G-PPP White Paper, 2021.

[79] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples.” arXiv preprint arXiv:1412.6572, 2014.

[80] Zujany Salazar, Huu Nghia Nguyen, Wissam Mallouli, Ana R. Cavalli, Edgardo Montes de

Oca: 5Greplay: a 5G Network Traffic Fuzzer Application to Attack Injection. ARES 2021: 106:1-106:8.

[81] Montimage security monitoring framework, https://www.montimage.com, as of date 29.03.2022

[82] Vinh Hoa La, Edgardo Montes de Oca, Wissam Mallouli, Ana R.Cavalli, "A Framework for Security Monitoring of Real IoT Testbeds." ICSOFT 2021: 645-652.

[83] EPC-in-a-Box: Allows The Instant Deployment Of A Core Network, https://www.montimage.com/products#EPC-in-a-box, as of date 29.03.2022

[84] B. Farahani, M. Barzegari, and F. S. Aliee, 'Towards Collaborative Machine Learning Driven Healthcare Internet of Things', in Proceedings of the International Conference on Omni-Layer Intelligent Systems, Crete Greece, May 2019, pp. 134–140. doi: 10.1145/3312614.3312644.

[85] J. B. Awotunde, S. O. Folorunso, A. K. Bhoi, P. O. Adebayo, and M. F. Ijaz, 'Disease Diagnosis System for IoT-Based Wearable Body Sensors with Machine Learning Algorithm', in Hybrid Artificial Intelligence and IoT in Healthcare, vol. 209, A. Kumar Bhoi, P. K. Mallick, M. Narayana Mohanty, and V. H. C. de Albuquerque, Eds. Singapore: Springer Singapore, 2021, pp. 201–222. doi: 10.1007/978-981-16-2972-3_10.

[86] W. Li et al., 'A Comprehensive Survey on Machine Learning-Based Big Data Analytics for IoT-Enabled Smart Healthcare System', Mob. Netw. Appl., vol. 26, no. 1, pp. 234–252, Feb. 2021, doi: 10.1007/s11036-020-01700-6.

[87] Y. Rebahi, K. T. Chiu, N. Tcholtchev, S. Hohberg, E. Pallis, and E. Markakis, 'Towards a next generation 112 testbed: The EMYNOS ESInet', Int. J. Crit. Infrastruct. Prot., vol. 22, pp. 39–50, Sep. 2018, doi: 10.1016/j.ijcip.2018.05.001.

[88] B. S. Kumar Subudhi, F. Catal, N. Tcholtchev, K. T. Chiu, and Y. Rebahi, 'Performance Testing for VoIP Emergency Services: a Case Study of the EMYNOS Platform', Procedia Comput. Sci., vol. 151, pp. 287–294, Jan. 2019, doi: 10.1016/j.procs.2019.04.041.

[89] R. Barakat, F. Catal, N. Tcholtchev, and Y. Rebahi, 'TTCN-3 based NG112 Test System and Playground for Emergency Communication', in 2020 IEEE 20th International Conference on Software Quality, Reliability and Security Companion (QRS-C), Dec. 2020, pp. 492–497. doi: 10.1109/QRS-C51114.2020.00088.

[90] E. Schooler et al., 'SIP: Session Initiation Protocol', Internet Engineering Task Force, Request for Comments RFC 3261, Jul. 2002. doi: 10.17487/RFC3261.

[91] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," Pattern Recognition, vol. 84, pp. 317– 331, 2018.

[92] TinyML: https://www.tinyml.org/ , as of date 30.03.2022

[93] Satyanarayanan, M., et al. (2021). The Role of Edge Offload for Hardware-Accelerated Mobile Devices. Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications: 22-29.

[94] Nvidia Jetson Nano: https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-nano, as of date 30.03.2022

[95] Katal, Avita, Mohammad Wazid, and Rayan H. Goudar. "Big data: issues, challenges, tools

and good practices." In 2013 Sixth international conference on contemporary computing (IC3), pp. 404-409. IEEE, 2013.

[96] Najafabadi, Maryam M., Flavio Villanustre, Taghi M. Khoshgoftaar, Naeem Seliya, Randall Wald, and Edin Muharemagic. "Deep learning applications and challenges in big data analytics." Journal of big data 2, no. 1 (2015): 1-21.

[97] Gudivada, Venkat, Amy Apon, and Junhua Ding. "Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations." International Journal on Advances in Software 10, no. 1 (2017): 1-20.

[98] Provost, Foster, and Tom Fawcett. "Data science and its relationship to big data and data-driven decision making." Big data 1, no. 1 (2013): 51-59.

[99] Shorten, Connor, and Taghi M. Khoshgoftaar. "A survey on image data augmentation for deep learning." Journal of big data 6, no. 1 (2019): 1-48.

[100] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[101] W. Stallings and M. P. Tahiliani, "Cryptography and network security: principles and practice, vol. 6," 2014.

[102] "Art. 4 GDPR – definitions," Mar 2018. [Online]. Available: https://gdpr-info.eu/art-4-gdpr/, as of date 30.03.2022

[103] R. Lu, H. Zhu, X. Liu, J. K. Liu and J. Shao, "Toward efficient and privacy-preserving computing in big data era," in IEEE Network, vol. 28, no. 4, pp. 46-50, July-August 2014, doi: 10.1109/MNET.2014.6863131.

[104] M. Liyanage, J. Salo, A. Braeken, T. Kumar, S. Seneviratne, and M. Ylianttila, "5g privacy: Scenarios and solutions," in 2018 IEEE 5G World Forum (5GWF). IEEE, 2018, pp. 197–203.

[105] Y. Sun, J. Liu, J. Wang, Y. Cao, and N. Kato, "When machine learning meets privacy in 6g: A survey," IEEE Communications Surveys & Tutorials, vol. 22, no. 4, pp. 2694–2724, 2020.

[106] A. Cavoukian, "Privacy by design," Identity in the Information Society, 2009.

[107] H. Zhang, Y. Shu, P. Cheng, and J. Chen, "Privacy and performance trade-off in cyber-physical systems," IEEE Network, vol. 30, no. 2, pp. 62–66, 2016.

[108] Horkoff, Jennifer. "Non-functional requirements for machine learning: Challenges and new directions." 2019 IEEE 27th International Requirements Engineering Conference (RE). IEEE, 2019.

[109] European Commission, Directorate-General for Communications Networks and Technology, "Artificial intelligence act," 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEXn%3A52021PC0206

[110] Vogelsang, Andreas, and Markus Borg. "Requirements engineering for machine learning: Perspectives from data scientists." 2019 IEEE 27th International Requirements Engineering Conference Workshops (REW). IEEE, 2019

[111] Mauritz Kop, "EU Artificial Intelligence Act: The European Approach to AI", online: https://law.stanford.edu/publications/eu-artificial-intelligence-act-the-european-

approach-to-ai/, as of date 30.03.2022

[112] Shaping Europe's digital future, online: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence, as of date 30.03.2022

[113] ANNEXES to the Proposal for a Regulation of the European Parliament and of the Council LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, online: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_2&format=PDF, as of date 30.03.2022

[114] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in AISTATS, ser. Proceedings of Machine Learning Research, vol. 108. PMLR, 2020, pp. 2938–2948.

[115] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. B. Calo, "Analyzing federated learning through an adversarial lens," in ICML, ser. Proceedings of Machine Learning Research, vol. 97. PMLR, 2019, pp. 634–643.

[116] C. Xie, K. Huang, P.-Y. Chen, and B. Li, "Dba: Distributed backdoor attacks against federated learning," in International Conference on Learning Representations, 2020. [Online]. Available: https://openreview:net/forum?id=rkgyS0VFvr

[117] C. Fung, C. J. M. Yoon, and I. Beschastnikh, "Mitigating sybils in federated learning poisoning," CoRR, vol. abs/1808.04866, 2018

[118] S. Li, Y. Cheng, W. Wang, Y. Liu, and T. Chen, "Learning to detect malicious clients for robust federated learning," CoRR, vol. Abs/2002.00211

[119] T. D. Nguyen, P. Rieger, H. Yalame, H. Möllering, H. Fereidooni, S. Marchal, M. Miettinen, A. Mirhoseini, A.-R. Sadeghi, T. Schneider, and S. Zeitouni, "FLGUARD: Secure and Private Federated Learning," 2021

[120] K. Pillutla, S. M. Kakade, and Z. Harchaoui, "Robust aggregation for federated learning," 2019

[121] S. Amershi et al., 'Guidelines for Human-AI Interaction', in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow Scotland Uk, May 2019, pp. 1–13. doi: 10.1145/3290605.3300233.

[122] J. Zhu, A. Liapis, S. Risi, R. Bidarra, and G. M. Youngblood, 'Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation', in 2018 IEEE Conference on Computational Intelligence and Games (CIG), Aug. 2018, pp. 1–8. doi: 10.1109/CIG.2018.8490433.

[123] A. Zytek, D. Liu, R. Vaithianathan, and K. Veeramachaneni, 'Sibyl: Understanding and Addressing the Usability Challenges of Machine Learning In High-Stakes Decision Making', IEEE Trans. Vis. Comput. Graph., vol. 28, no. 1, pp. 1161–1171, Jan. 2022, doi: 10.1109/TVCG.2021.3114864.

[124] S. Oesch et al., 'An Assessment of the Usability of Machine Learning Based Tools for the Security Operations Center', ArXiv201209013 Cs, Dec. 2020, Accessed: Mar. 12, 2022. [Online]. Available: http://arxiv.org/abs/2012.09013

[125] Q. V. Liao, D. Gruen, and S. Miller, 'Questioning the AI: Informing Design Practices for Explainable AI User Experiences', in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu HI USA, Apr. 2020, pp. 1–15. doi: 10.1145/3313831.3376590.

[126] Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021, March). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 624-635).

[127] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 2125–2126. DOI:https://doi.org/10.1145/2939672.2945386

[128] "A European approach to artificial intelligence | Shaping Europe's digital future." https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence, as of date 31.03.2022

[129] Kubeflow - https://www.kubeflow.org/, as of date 01.04.2022

[130] Enarx - https://enarx.dev/, as of date 01.04.2022

[131] Veracruz - https://veracruz-project.com/, as of date 01.04.2022

[132] Mainflux Edge - https://docs.mainflux.io/edge/, as of date 01.04.2022

[133] Edgex SMA - https://docs.edgexfoundry.org/1.2/microservices/system-management/agent/Ch_SysMgmtAgent/, as of date 01.04.2022

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

APPENDIX A: SUMMARY OF IDENTIFIED REQUIREMENTS

In the following section, we provide a short and precise summary of the identified and specified requirements that r.epresent aspects and general design principles to be considered when integrating and utilizing AI algorithms and frameworks for addressing security risks and threats to system (and network) architectures.

A.1 SOFTWARE REQUIREMENTS

TABLE 2: SUMMARY OF GATHERED SOFTWARE REQUIREMENTS

Identifier	Software Requirements	Functional/ Non-functional	Priority	Relevant for	Implemented by
SW.RQ.1	Microservices developed in the context of the SPATIAL project SHOULD be written as cloud native apps with Kubernetes as a default deployment orchestrator.	Functional	SHOULD	SPATIAL Use Cases, SPATIAL Platform Components	Platform, UC1, UC2, UC3, UC4
SW.RQ.2	Kubernetes orchestrator utilized in the SPATIAL project SHOULD have capabilities to enable federated AI execution.	Functional	SHOULD	SPATIAL Use Cases, SPATIAL Platform Components	Platform, UC1
SW.RQ.3	AI-based systems deployed at the Edge or in the Cloud SHOULD use software enablers for Trusted Execution Environments in order to enable secure and confidential computations.	Functional	SHOULD	In General	

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

SW.RQ.4	Edge nodes SHOULD use Software Management Agents that are capable monitoring and assessing processes as well as to exchange messages with the cloud control software for various purposes.	Functional	SHOULD	In General	
SW.RQ.5	The Software Management Agents deployed to edge nodes SHOULD be capable of receiving commands from the cloud and executing them in order to control other processes.	Functional	SHOULD	In General	

A.2 HARDWARE REQUIREMENTS

TABLE 3: SUMMARY OF IDENTIFIED HARDWARE REQUIREMENTS

Identifier	Hardware Requirements	Functional/ Non-functional	Priority	Relevant for	Implemented by
HW.RQ.1	An AI-based system in the scope of the SPATIAL use cases SHOULD include a distributed edge infrastructure.	Functional	SHOULD	SPATIAL Use Cases	UC1, UC2
HW.RQ.2	An AI-based system deployed at the edge MAY support hardware acceleration at the edge.	Functional	MAY	In General, SPATIAL Use Cases	UC2

SPATIAL project is funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 101021808.

HW.RQ.3	An AI-based system in the scope of the SPATIAL use cases SHOULD be able to operate in a 4G/5G network setting.	Functional	SHOULD	SPATIAL Use Cases	UC2
HW.RQ.4	An AI-based system in the scope of the SPATIAL use cases SHOULD be able to operate in an IoT network setting including on small scale and resource constrained devices.	Functional	SHOULD	SPATIAL Use Cases	UC1, UC2
HW.RQ.5	In the context of IoT networks, an AI-based system MAY include constrained nodes.	Functional	MAY	SPATIAL Use Cases	UC1, UC2
HW.RQ.6	In the context of Use Case 3, an AI-based system SHOULD be able to operate on VoIP enabled end devices.	Functional	SHOULD	SPATIAL Use Cases	UC3
HW.RQ.7	The AI-based system reflected in Use Case 3 SHOULD be able to operate on health sensors.	Functional	SHOULD	SPATIAL Use Cases	UC3
HW.RQ.8	In the scope of Use Case 3, the AI-based system SHOULD be able to utilize information from a location information server.	Functional	SHOULD	SPATIAL Use Cases	UC3

A.3 DATA REQUIREMENTS

TABLE 4: SUMMARY OF IDENTIFIED DATA REQUIREMENTS.

Identifier	Data Requirements	Functional/ Non-functional	Priority	Relevant for	Implemented by
DAT.RQ.1	The data used to train, optimize, and validate AI models MUST be a good representative of the use case for which the models will be applied in practice.	Non-functional	MUST	In General, SPATIAL Use Cases	UC1, UC2, UC3, UC4
DAT.RQ.2	The features and labels in the supplied data MUST follow a similar distribution as the data encountered in the production environment.	Non-functional	MUST	In General, SPATIAL Use Cases	UC2, UC4
DAT.RQ.3	The training, validation, and testing data SHOULD contain a sufficient number of the edge and corner cases that could potentially occur in practice.	Non-functional	SHOULD	In General, SPATIAL Use Cases	UC1, UC2, UC3
DAT.RQ.4	The data set MUST be fair and unbiased so that discrimination in the model predictions based on sensitive inputs (e.g. ethnicity, health, gender, religion, race etc.) can be and MUST be prevented at all costs.	Non-functional	MUST	In General, SPATIAL Use Cases	UC3

SPATIAL project is funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 101021808.

DAT.RQ.5	Data preparation and management processes MUST be adopted to easily replicate and train AI models from raw data.	Non-functional	MUST	In General, SPATIAL Use Cases	UC1, UC2, UC3, UC4
DAT.RQ.6	Heterogeneous data aggregation MUST rely on automatic procedures, which resolve data issues related with duplication, inconsistencies, and missing data.	Functional	MUST	In General, SPATIAL Use Cases	UC2
DAT.RQ.7	Before training a model, private and sensitive data MUST be removed from the dataset.	Non-functional	MUST	In General, SPATIAL Use Cases	UC3
DAT.RQ.8	Feature/variable extraction for defining input data formats for models SHOULD analyse data automatically for correlation, outlier removal, and data transformations.	Functional	SHOULD	In General, SPATIAL Use Cases	UC2
DAT.RQ.9	After data is prepared, the pre-processed data SHOULD be analysed to ensure that data semantics are preserved.	Non-functional	SHOULD	In General, SPATIAL Use Cases	UC2
DAT.RQ.10	Data quality SHOULD be measured by quantifying the performance of the AI model.	Non-functional	SHOULD	In General, SPATIAL Use Cases	UC2

DAT.RQ.11	Data quality for AI training SHOULD be also explicitly defined by data dimensions (e.g. accuracy, currency, and consistency).	Non-functional	SHOULD	In General, SPATIAL Use Cases	UC2
DAT.RQ.12	Pre-processed data MAY be enriched further before training AI models to improve robustness and performance	Non-functional	MAY	In General, SPATIAL Use Cases	UC2
DAT.RQ.13	Pre-processed input data SHOULD be linked with prediction outputs of AI models to derive quantifiable explanations to users.	Functional	SHOULD	In General, SPATIAL Use Cases, SPATIAL Platform Components	Platform, UC2, UC3
DAT.RQ.14	AI-models can be continually trained with aggregated data, but consistency and integrity of data MUST be preserved through quantifiable estimations.	Non-functional	MUST	In General, SPATIAL Use Cases	UC1

A.4 DATA PRIVACY REQUIREMENTS

TABLE 5: SUMMARY OF CAPTURED DATA PRIVACY REQUIREMENTS.

Identifier	Data Privacy Requirements	Functional/ Non-functional	Priority	Relevant for	Implemented by
PRV.RQ.1	The category of collected data SHOULD be identified for maintaining privacy measures based on the category.	Non-functional	SHOULD	In General, SPATIAL Use Cases	UC1
PRV.RQ.2	Privacy measures for data SHOULD be considered in each stage of data generation, processing, and storage.	Non-functional	SHOULD	In General, SPATIAL Use Cases	UC1
PRV.RQ.3	The developments MAY maintain data portability when preserving privacy.	Non-functional	MAY	In General	
PRV.RQ.4	The system MUST be able to make updates or erasure of personal data if required by data owners.	Functional	MUST	In General	
PRV.RQ.5	The possibility of privacy-related attacks on AI, system and data MUST be assessed and protection or mitigation processes MUST be made.	Non-functional	MUST	In General, SPATIAL Use Cases	UC1

PRV.RQ.6	There SHOULD be proper metrics defined for privacy to support privacy protection measures	Non-functional	SHOULD	SPATIAL Platform Components	Platform
PRV.RQ.7	The privacy by design approaches SHOULD be included during the system design process.	Non-functional	SHOULD	In General, SPATIAL Use Cases	UC1
PRV.RQ.8	Trade-offs between model performance and privacy MAY be considered when implementing privacy.	Non-functional	MAY	In General, SPATIAL Use Cases	UC1

A.5 MODEL REQUIREMENTS

TABLE 6: SUMMARY OF IDENTIFIED MODEL REQUIREMENTS

Identifier	Model Requirements	Functional/ Non-functional	Priority	Relevant for	Implemented by
MOD.RQ.1	The ML model MUST have a high accuracy.	Non-functional	MUST	In General, SPATIAL Use Cases	UC1, UC2, UC3, UC4

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

MOD.RQ.2	The high performance of the ML model MUST generalize to unknown data.	Non-functional	MUST	In General, SPATIAL Use Cases	UC1, UC2, UC3, UC4
MOD.RQ.3	With respect to maintainability, ML models MUST adapt to changes in their environment when these changes occur.	Functional	MUST	In General, SPATIAL Use Cases	UC1, UC4
MOD.RQ.4	ML models SHOULD provide their predictions quickly with short response time and latency.	Non-functional	SHOULD	In General, SPATIAL Use Cases	UC1, UC2, UC3
MOD.RQ.5	To guarantee fairness, ML models SHOULD NOT output biased decisions.	Non-functional	SHOULD NOT	In General, SPATIAL Platform Components	Platform
MOD.RQ.6	ML models' predictions SHOULD provide high-level of explainability and should be understandable by humans.	Non-functional	SHOULD	In General, SPATIAL Use Cases, SPATIAL Platform Components	Platform, UC2, UC3
MOD.RQ.7	The training, deployment and usage of ML models MUST be documented and accessible, in order to provide high-level of transparency.	Non-functional	MUST	In General, SPATIAL Use Cases	UC2, UC3

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

MOD.RQ.8	ML models SHOULD be testable to verify they fulfil expectations on their outputs.	Non-functional	SHOULD	In General, SPATIAL Use Cases	UC1, UC2, UC4
MOD.RQ.9	ML models SHOULD be trained in a reasonable amount of time and be scalable with respect to the amount of data, devices, and services.	Non-functional	SHOULD	In General, SPATIAL Use Cases	UC1, UC2, UC4

A.6 LEGISLATIVE REQUIREMENTS

TABLE 7: SUMMARY OF IDENTIFIED LEGISLATIVE REQUIREMENTS

Identifier	Legislative Requirements	Functional/ Non-functional	Priority	Relevant for	Implemented by
LEG.RQ.1	According to Article 12 of the GDPR, there MUST be a process to provide information to the data subject in an easily accessible form, in which they will be able to comprehend.	Non-functional	MUST	In General	
LEG.RQ.2	According to article 15 of the GDPR, there MUST be a process to provide the data subject's data upon request within reasonable time (1 Month).	Non-functional	MUST	In General	

LEG.RQ.3	According to article 21 of the GDPR, there MUST be a process to allow data subjects to object to some form of data processing of any data concerning them.	Non-functional	MUST	In General	
LEG.RQ.4	According to article 22 of the GDPR, there MUST be a process to ensure that fully automated decisions resulting in any legal or other significant effects are not taken without the intervention of a human.	Non-functional	MUST	In General	
LEG.RQ.5	According to the AI Act, AI owners/developers MAY need to create and maintain a risk management system for the entire lifecycle of the (AI-based) system.	Non-functional	MAY	In General	
LEG.RQ.6	According to the AI Act, there MAY need to be a testing process to identify risks and determine appropriate mitigation measures, and to validate that the system runs consistently for the intended purpose, with tests made against prior metrics and validated against probabilistic thresholds.	Non-functional	MAY	In General, SPATIAL Use Cases	UC4
LEG.RQ.7	According to the AI Act, appropriate data governance controls MAY need to be established, including the requirement that all training, validation, and testing datasets be complete, error-free, and representative	Non-functional	MAY	In General	

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

LEG.RQ.8	According to the AI Act, there MAY need to be detailed documentation on the AI, including around system architecture, algorithmic design, and model specifications.	Non-functional	MAY	In General, SPATIAL Use Cases	UC1, UC2, UC3
LEG.RQ.9	According to the AI Act, AI systems MAY need automatic logging of events while the system is running, with recording conforming to recognized standards.	Functional	MAY	In General	
LEG.RQ.10	According to the AI Act, AI MAY need to be designed with sufficient transparency to allow users to interpret the system's output.	Non-functional	MAY	In General, SPATIAL Use Cases, SPATIAL Platform Components	Platform, UC2, UC3
LEG.RQ.11	According to the AI Act, AI systems MAY need to be designed to maintain human oversight at all times and prevent or minimize risks to health and safety or fundamental rights, including an override or off-switch capability.	Non-functional	MAY	In General, SPATIAL Use Cases	UC4

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

A.7 SECURITY REQUIREMENTS

TABLE 8: SUMMARY OF IDENTIFIED SECURITY REQUIREMENTS

Identifier	Security Requirements	Functional/ Non-functional	Priority	Relevant for	Implemented by
SEC.RQ.1	AI-based systems MUST be resilient against the evasion attacks.	Non-functional	MUST	In General, SPATIAL Use Cases	UC2, UC4
SEC.RQ.2	AI-based systems MUST be resilient against data poisoning attacks.	Non-functional	MUST	In General, SPATIAL Use Cases	UC1, UC2
SEC.RQ.3	AI-based systems MAY apply anti-poisoning techniques (e.g., data sanitization) to training data obtained from untrusted sources.	Functional	MAY	In General, SPATIAL Use Cases	UC1, UC2
SEC.RQ.4	AI-based systems MUST be resilient against backdoor attacks.	Non-functional	MUST	In General	
SEC.RQ.5	AI-based systems MUST be resilient against model extraction attacks.	Non-functional	MUST	In General	
SEC.RQ.6	AI-based systems MUST be resilient against data reconstruction attacks.	Non-functional	MUST	In General, SPATIAL Use Cases	UC1

SPATIAL project is funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 101021808.

SEC.RQ.7	AI-based systems MUST be resilient against property interference attacks.	Non-functional	MUST	In General	
SEC.RQ.8	AI-based systems MUST be resilient against membership interference attacks.	Non-functional	MUST	In General	
SEC.RQ.9	AI-based systems that apply online learning strategies SHOULD be resilient against attacks that manipulate supplied data and aim to attack the system through manipulated retraining of the underlying AI model.	Non-functional	SHOULD	In General	
SEC.RQ.10	AI-based systems SHOULD be resilient against DoS and DDoS attacks.	Non-functional	SHOULD	In General, SPATIAL Use Cases	UC2
SEC.RQ.11	AI-based systems, dealing with sensitive or confidential data, MUST preserve the confidentiality of the data during the operational phase.	Non-functional	MUST	In General	

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

A.8 USABILITY REQUIREMENTS

TABLE 9: SUMMARY OF IDENTIFIED USABILITY REQUIREMENTS.

Identifier	Usability Requirements	Functional/ Non-functional	Priority	Relevant for	Implemented by
USB.RQ.1	AI-based systems MUST provide comprehensible, uniform, and easy-to-use interfaces.	Non-functional	MUST	In General, SPATIAL Use Cases, SPATIAL Platform Components	Platform, UC1, UC2, UC3
USB.RQ.2	AI-based systems MUST provide detailed and comprehensive documentation that should include extensive information about the data used, the training process, as well as information about the utilized AI models and their deployment.	Non-functional	MUST	In General, SPATIAL Use Cases	UC1, UC2, UC3
USB.RQ.3	An AI-based system SHOULD have functionalities that guide users in the usage of the system and provide help in case of problems.	Functional	SHOULD	In General, SPATIAL Use Cases	U2
USB.RQ.4	All decisions and outputs of AI-based systems SHOULD be as consistent as possible and follow pre-specified and interpretable formats.	Non-functional	SHOULD	In General, SPATIAL Use Cases, SPATIAL Platform Components	Platform, UC2, UC3, UC4

SPATIAL project is funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 101021808.

USB.RQ.5	All user interfaces SHOULD be consistent and unified in order to increase the usability of the system.	Non-functional	SHOULD	In General, SPATIAL Use Cases, SPATIAL Platform Components	Platform, UC2, UC4
USB.RQ.6	AI-based systems MUST provide explanations for individual decisions of the deployed AI models.	Functional	MUST	In General, SPATIAL Use Cases, SPATIAL Platform Components	Platform, UC2, UC3
USB.RQ.7	An AI-based system SHOULD provide information about the effect of concrete AI decisions to the users.	Functional	SHOULD	In General	
USB.RQ.8	AI-based system MAY provide functionalities that offer support to the users in case of system errors.	Functional	MAY	In General , SPATIAL Use Cases	UC2
USB.RQ.9	Users of AI-based systems SHOULD be able to identify, report, and correct mistakes in the decision-making of AI models.	Functional	SHOULD	In General , SPATIAL Use Cases, SPATIAL Platform Components	Platform, UC2, UC4
USB.RQ.10	Users of AI-based systems MUST be informed about every system update.	Functional	MUST	In General, SPATIAL Use Cases	UC2

USB.RQ.11	Only minor and incremental system updates and adaptions SHOULD be applied to AI-based systems.	Non-functional	SHOULD	In General	

A.9 ACCESSIBILITY REQUIREMENTS

TABLE 10: SUMMARY OF ACCESSIBILITY REQUIREMENTS

Identifier	Accessibility Requirements	Functional/ Non-functional	Priority	Relevant for	Implemented by
ACC.RQ.1	The purpose of the ML solution MUST be specified in the documentation	Non-functional	MUST	In General, SPATIAL Use Cases, SPATIAL Platform Components	Platform, UC1, UC2, UC3, UC4
ACC.RQ.2	The inherent risks of the ML solution, including the potential for bias and discrimination, SHOULD be listed in the documentation	Non-functional	SHOULD	In General	
ACC.RQ.3	Non-Functional requirements relating to the ML process (e.g. accuracy and generalizability) SHOULD be documented in a way that is accessible to lay users	Non-functional	SHOULD	In General, SPATIAL Use Cases, SPATIAL Platform Components	Platform, UC1, UC2, UC3

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

ACC.RQ.4	Documentation MUST specify who is accountable for each component (data, model, outputs) of the ML solution	Non-functional	MUST	In General	

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

APPENDIX B: SHORT OVERVIEW OF EXPLAINABLE AI METHODS

In recent years, many different approaches emerged that aim to provide explanations of the behaviour of AI algorithms and hence increase their explainability. Since many of these approaches exhibit significantly different characteristics, we will present a brief insight into the taxonomy of explainable AI approaches in the following section. Afterward, we will briefly introduce some relevant state-of-the-art XAI methods that can support the explainability and thus the accountability of AI-based systems and networks.

B.1 XAI TAXONOMY

The XAI approaches may be divided into multiple categories based on various criteria [48] [49]. The most common XAI based taxonomies are discussed below. XAI methods that fall into those categories are not necessarily exclusive for each group. There can be methods that belongs to even two or more categories according to the taxonomy.

Model-agnostic vs Model-specific methods

Model-agnostic methods for XAI are the ones that are not constrained by the core parts of an AI algorithm when making a prediction. They are useful in decoding the decision process of black box models and provide good flexibility for developers to apply it to a wide variety of ML models. On the contrary, model specific methods are bespoken for specific models and take the use of core components of an ML model to interpret the outcomes. This makes model-specific methods more suitable to identify granular aspects of ML models but they lack the flexibility.

Local vs Global methods

XAI methods can be divided into two main categories based on the scope of the function that's used by the interpreter to produce an explanation. They are local and global explanations. Local explainers are designed to interpret a portion of the model function that contributes towards the outcome given a certain datapoint. The close vicinity of ML function to that datapoint is explored when generating an explanation. On the contrary, global methods take the ML function as a whole when generating explanations for the inference. This generally makes these methods slow but robust where local methods are fast but erratic sometimes.

Pre-model, In-model, and Post-model explainers

Depending on the stage at which XAI methods are applied in the development process, there are three main categories of XAI: pre-model, in-model, and post-model. Pre-model methods are mainly used during the dataset preparation time in the model development pipeline. These methods are useful in data analysis, feature engineering, and explaining any underlying patterns seen in data at a glance. In-model XAI methods are embedded in the ML algorithms. This includes all the transparent models such as linear regression or decision trees. In addition, in-model

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

explanations are also generated through modifications done to the existing ML model architectures using inherently transparent models. Post-hoc/post-model explanations are applied after training an ML model. This enables us to identify what the model has learned during the training process.

Surrogate vs Visualization

XAI method are divided into two main categories based on what exactly is explained during the process. Surrogate model-based explainers generate explanations from an approximated model of the black-box model that is trained in a similar way to mimic the original model's behaviour. Surrogate models are mostly inherently interpretable. Otherwise, one can use visualizations techniques (e.g.: heatmaps, graphs, etc.) on the original black-box model to explore the internal workings of them without using a representation. These XAI methods fall under the visualization category.

B.2 XAI METHODS

This section briefly introduces some relevant XAI methods. To not exceed the deliverable at hand, we refer to yet-to-come deliverables of the SPATIAL project for a detailed discussion of the presented approaches.

B.2.1 LIME

LIME [18] is a widely popular technique used in interpreting outputs of black-box models in several fields and applications. LIME is short for Local Interpretable Model-agnostic Explanations. As the name suggests, LIME gives a *local* explanation, which means that it considers a subset of data when approximating explanations for model predictions. This technique is plausible under the premise that every complicated model performs linearly on a local scale. Nevertheless, LIME has recently gained a reputation owing to its speed (relative to global explanation techniques) and convenience as it can interpret outputs irrespective of the type of black-box model (*model-agnostic)* which it wraps around.

B.2.2 SHAP

SHAP (Shapley Additive Explanations) [50] is a *model-agnostic* XAI technique that identifies the importance of each feature value in a certain prediction. For explaining individual predictions, it uses a concept called Shapley values. These Shapley values are a popular cooperative game theory technique that is based on the question of distributing a reward fairly among players of a group. Since the contribution of players for the winning could be different, the reward should also be based on it. This concept is applied in order to explain *local* AI predictions, to identify how features are contributing different amount to the final prediction. For this, Shapley values are used to calculate the contribution of each feature to the prediction by determining its marginal contribution for each possible set of features.

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

B.2.3 Counterfactual Explanations

Counterfactual explanations are a *local model-agnostic* XAI method that explains an AI model prediction/decision by answering the question "How should my input X be different in order to achieve my desired outcome Y?" [51]. Therefore, the main methodology behind counterfactual explanations is to examine a hypothetical scenario that is opposite to what is currently observed [52]. Figure 8 illustrates visually this idea. What the diagram shows is a hypothetical use case where a person is trying to apply for a bank loan and the application was rejected. As the figure suggests we can generate two counterfactual explanations such as: 1.) "had you increased your income with $10,000, you would have received the loan" or 2.) "had you increased your income with $5,000 and improved your credit score, you would have received the loan". Both of these explanations show how the input features (i.e. annual income and credit score) have to be different than the current observed feature values, in order for the model to generate the desired output (i.e. approved loan).

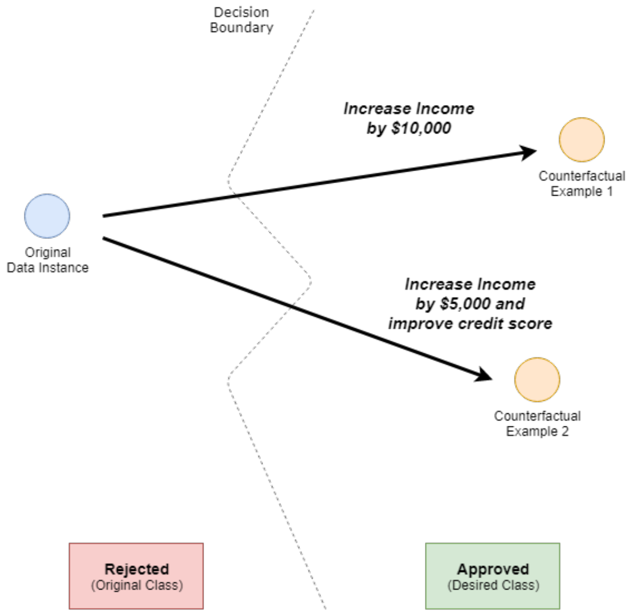

FIGURE 8: EXAMPLE FOR TWO COUNTERFACTUAL EXPLANATIONS FOR A LOAN APPLICATION (ADAPTED FROM [53])

This example illustrates the main advantage of counterfactual explanations – they provide human-friendly explanations and focus only on a selected subset of features of interest [52], which makes it easier to interpret how select few causes could lead to a specific desired outcome [52]. However, counterfactual explanations could suffer from the so-called "Rashomon effect", which stems from the fact that typically there is more than one counterfactual explanation for a single instance (see Figure 8). In some cases, counterfactual explanations for the same instance could even contradict each other [51], which makes it difficult to determine which explanation is "good" and which is "bad".

B.2.4 Permutation Feature Importance

When it comes to training an accurate Machine Learning model, having high quality input data and features with high predictive power is of crucial importance. In order to evaluate the predictive power of each individual input feature, the so-called "feature importance" score can be generated. One *global* and *model-agnostic* XAI method that allows measuring the feature importance score presented by L. Breiman [54] is ***permutation feature importance***. Permutation feature importance is a technique that computes the predictive power of each feature by randomly shuffling its values and observing the effects that this shuffling procedure will have on the prediction error [52]. The main intuition here is that changing (or shuffling) the values of the less "important" features should not affect the prediction error in any significant way. Conversely, since the model relies heavily on the "important" features to generate its predictions, shuffling their values would result in a considerably increased prediction error.

B.2.5 T-SNE

T-distributed Stochastic Neighbor Embeddings (t-SNE) is a non-linear, unsupervised statistical tool for dimensionality reduction [55]. It maps high dimensional data into an alternative low-dimensional representation (typically a two or three dimensional space [56]), which makes visual interpretation of these data points easier. At the same time, the method aims at minimizing the information loss between the high and low dimensional data representation and consequently it tries to preserve most of the high dimensional data patterns. In this way, the ML practitioner could gain a better intuition about the arrangement of the data points in the high-dimensional space, which is why t-SNE can be used as a pre-model XAI method. The main idea behind t-SNE is illustrated in Figure 9, where points from a 2D space (i.e. high dimension) are mapped to their equivalent low-dimensional (i.e. 1D) representation. In simple terms, t-SNE works in two steps. First, the algorithm computes the similarity between points in the high-dimensional space by generating a probability distribution P, such that if a given data point "x_i" is very similar to another point "x_j", the probability $P(x_j|x_i)$ will also be very high. The second step of t-SNE is to generate another probability distribution Q in the low-dimensional space, such that Q is as similar as possible to P. These steps are based on another algorithm called Stochastic Neighbor Embedding (SNE) introduced by G. Hinton and S. Roweis [56]. However, t-SNE improves upon the standard SNE by utilising a more better cost function and an alternative distribution for the similarity comparison between points in low dimensional space [55].

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

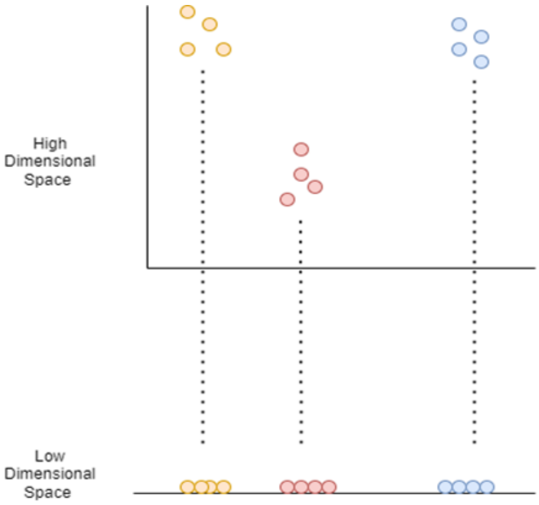

FIGURE 9: VISUAL EXAMPLE OF T-SNE

With this in mind, among the advantages of t-SNE is that it is well-suited for handling non-linear data and it manages to preserve the local and global structure of the data [55]. However, t-SNE is computationally expensive to run, might require hyperparameter optimization, and it is not always trivial to interpret the algorithm results [57].

B.2.6 Layer-wise Relevance Propagation

Layer-Wise Relevance Propagation (LRP) is a *local model-specific* XAI method that explains Neural Network predictions with respect to their inputs [58]. The method shows which input features contributed most for the model decision (see heatmap in the lower network in Figure 10). The main idea of LRP is to use the weights and activations computed during the forward propagation pass through a Neural Network (NN) to calculate the so-called neuron "relevance". As the name suggests, the relevance score is a measure of how relevant a particular neuron is for the prediction generated by the model. The relevance is computed with a backpropagation from the output layer back the input layer (see lower part in Figure 10) with the help of the so-called "propagation rules". The figure illustrates the propagation of relevance scores from the output back to the input layer, where darker colours indicate higher neuron relevance. Once the relevance scores are backpropagated to the input layer, we can generate a heatmap (see Figure 10) which indicates which features contributed most for the classification decision made by the model.

Important to note is that LRP is a technique that operates according to the "conservation" property, which states that the relevance values computed in the output layer are redistributed to the lower layers without any losses [58]. In other words, the output layer relevance scores are preserved throughout the backpropagation process back to the input layer [58]. LRP is particularly useful in safety-critical domains where knowing what the model does is of crucial importance. Among the main advantages of the method is that it provides human-understandable explanations. It can be implemented efficiently and is very flexible due to the wide range of different LRP propagation rules.

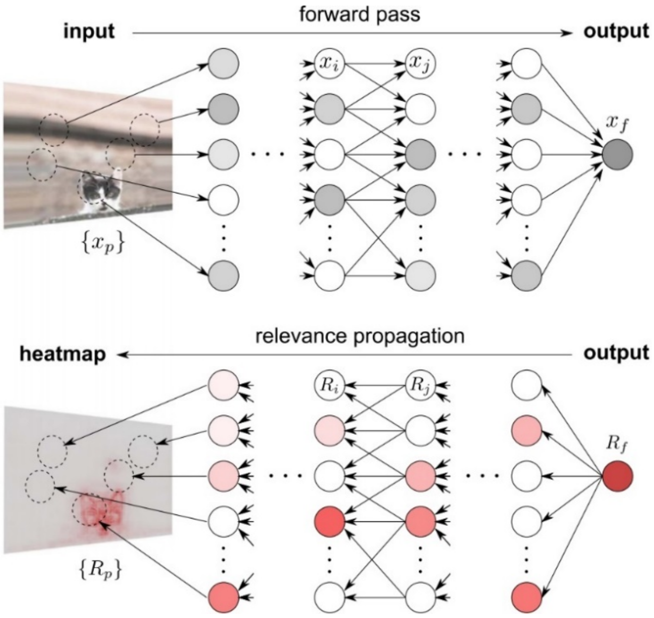

FIGURE 10: BACKPROPAGATION OF RELEVANCE SCORES BACK TO THE INPUT LAYER TAKEN FROM [59]

B.2.7 Occlusion Sensitivity

Occlusion sensitivity is a *model-agnostic*, and a visual-based explainability method. It generates *local explanations* using maps independently of the underlying model by covering the pixel of input data (in case of an image) with an occlusion mask or patch, for model class prediction. The variation in the model's prediction of the occluded input image from the original input image is captured visually through saliency (feature) map or heatmap and numerically using metrics of feature relevance. The explanation of the inference of the model lies in the comparison of observed differences in probability of the predicted class of original input and the masked input measured using techniques like: Sensitivity-n, Spearman rank correlation co-efficient, Top-k intersection and Local Lipschitz continuity [60].

Zeiler & Fergus (2013) occlusion sensitivity experiment [61] for spatial understanding provides empirical evidence for the occlusion-based explanation by systematically occluding different portions of the input image with grey square (occlusion mask) and observing the drop accuracy of the classifier output [61]. This implementation showed that Convolutional Neural Networks (CNNs) locate the object in the images, and not the surrounding context of the object when making predictions. Also, the implementation provided intuitions on the logic of the model by revealing activated relevant features of the images that the model considered for its prediction.

Occlusion sensitivity steps:

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

- **Step 1:** Run trained model with original input for initial label or class and probability of class
- **Step 2:** Observe the class probability.
- **Step 3:** Occlude pixels with an occluding mask, i.e. a small grey square patch
- **Step 4:** Rerun the model inference on the occluded images.
- **Step 5:** Observe the changes in the probability of prediction class
- **Step 6:** Generate heatmap to observe the region in the heatmap that the model considered during inference generation.

B.2.8 CAM and Grad-CAM

Class Activation Mapping (CAM) and Gradient-weighted CAM (Grad-CAM) are both *local model-specific* XAI methods that use back propagation for interpretations, and are applicable to image data. This means that they interpret the models by providing the important features in a given image which causes the corresponding outcome [63]. CAM is used to interpret a special type of CNN, which uses the global average pooling (GAP) layer just before the final output layer. It was observed that using GAP in this way helps to avoid overfitting and was used to regularize training data [64]. In CAM, it was observed that using the weights from the last layer and calculating the weighted average of the feature map generated the localization map of features, which caused the output [62]. Thus, CAM can localize and highlight the discriminative parts of an object in a given image which correspond to the output class provided by the CNN.

In contrast to CAM, which is applicable to a particular CNN network with GAP layer, Grad-CAM can be applied for any CNN network. It generalizes CAM by taking the gradient of the scores for a given class with respect to the feature map activations, and then performs the global average pooling. After this, it performs the weighted average using these new weights, which is similar to CAM. In Grad-CAM, an additional Rectified Linear Unit (ReLU) layer is added after this to generate the feature visualizations. Thus, the steps before ReLU provide a generalization of the CAM algorithm so that it can also be applied to other fully connected CNN networks [65] [66].

B.2.9 Partial Dependence Plot

A partial dependence plot (PDP) [67] is an XAI method used for the visualization of the dependencies between an input feature and the prediction generated by an AI model. The method works by modifying the value of a selected feature for all samples in the data set, while maintaining the same values for the rest of the input features. Then, the dependency between the feature and the prediction is visualized with the help of a 2D or 3D plot. In terms of implementation, the algorithm is intuitive, straightforward to implement and fairly easy to interpret [52]. However, one major downside of PDP is that it operates under the assumption that the selected feature(s) do not correlate with the remaining features [52]. In practice, this assumption rarely holds true. Additionally, as pointed out by C. Molnar [52], since PDP provides explainability by means of visualization (i.e. in at most a 3-dimensional space), the number of features that can be selected in a partial dependence function is at most two [52]. This is another constraint when applying PDP.

SPATIAL project is funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 101021808.