6GSNS

Co-funded by the European Union

SAFE-6G

SAFE-6G

A Smart and Adaptive Framework for Enhancing Trust in 6G Networks

Deliverable D2.2: Overall SAFE-6G Framework and Reference Architecture Design

Date: 18/12/2025

Version: v1.1

DISCLAIMER

Grant Agreement	101139031
Document number	D2.2
Document title	Overall SAFE-6G Framework and Reference Architecture Design
Lead Beneficiary	NCSRD
Editor(s)	Harilaos Koumaras, Ilias Alexandropoulos, Vasiliki Rentoula, Spyridon Georgoulas (NCSRD)
Author(s)	Daniel García (TID) Andrés Anaya (TID) Javier García (TID) Ilias Alexandropoulos (NCSRD) Spyridon Georgoulas (NCSRD) Vasiliki Rentoula (NCSRD) Harilaos Koumaras (NCSRD) Zouzias Dimitris (eBOS) Marios Sophocleous (eBOS) Alejandro Fornés (UPV) Francisco Mahedero (UPV) Joaquín Cáceres (ATOS) Sonia Castro (ATOS) Ricardo Marco (ATOS) Apostolos Garos (SHG) Victoria Katsarou (SHG) Nikolaos Vryonis (INF) Eugenia Vergi (INF) George Koumaras (INF) Gaëtan Pruvost (THA) Alan Branchereau (THA) Stéphane Lorin (THA) Charles Bailly (IMM) Christos Xenakis (IQBT) Kushal Mehta (IQBT) Panos Karkazis (UNIWA) Dimitris Uzunidis (UNIWA) Stamatia Drampalou (UNIWA) Jose Costa-Requena (CMC) Nikolaos Zombakis (8BELLS) Guillaume Hébert (KEY) Van Hoan Hoang (KEY)
Dissemination level	Public
Contractual date of delivery	31/10/2024
Status	Final
File name	SAFE-6G_D2.2_V1.1.pdf

Revision History

Version	
V0.1	Allocation of work per partner
V0.2	Table of Contents
V0.3	Initial Contribution by all partners
V0.4	Second round of contributions by all partners
V0.5	First revision and homogenization of the content conducted by CMC and THA
V0.6	Additional input in cognitive coordination section
V0.7	Additional input related to MLOps and XAI
V0.8	Additional input related to data view section
V0.9	Final draft and homogenization of content produced by NCSRD based on the comments from the Second Review performed by the Technical Steering Committee.
V1.0	Final version following the Quality check
V1.1	Updated version after mid-term review

GLOSSARY

Abbreviations/Acronym	Description
AD	Architecture Descriptions
AF	Application Functions
AI	Artificial Intelligence
AKA	Authentication and Key Agreement
ALE	Accumulated Local Effects
AMF	Access Management Function
AnA	Authentication and Authorization
AP	Access Point
API	Application Programming Interface
AR	Augmented Reality
B5G	Beyond 5G
BSS	Business Support Systems
CAI	Conversational AI
CDP	Central Differential Privacy
CN	Core Network
CNCF	Cloud Native Computing Foundation
CNFs	Cloud Native Network Functions
CNI	Container Network Interface
CNNs	Convolutional Neural Networks
CTI	Cyber Threat Intelligence
DAGs	Direct Acyclic Graphs
DID	Decentralized Identifier
DLT	Distributed Ledger Infrastructure
DMS	Distribution Management Systems
DP	Differential Privacy
DSS	Decision Support System
DT	Digital Twin
E2E	End-to-End
EAP	Extensible Authentication Protocol
EC	European Commission
ENISA	European Union Agency for Cybersecurity
EU	European Union
FaaS	Function-as-a-Service
FL	Federated Learning
FLCM	Function Lifecycle Management
FLISR	Fault Location, Isolation, Service Restoration
GANs	Generative Adversarial Networks
GDPR	General Data Protection Regulation
HA	High Availability
HLO	High-Level Orchestrator
IaaS	Infrastructure-as-a-service
IBI	Intent-based Interface
IDS	Intrusion Detection System
IEs	Infrastructure Elements
InfP	Infrastructure Provider

IoT	Internet of Things
IPS	Intrusion Prevention System
ISMS	Information Security Management Systems
LDP	Local Differential Privacy
LLM	Large Language Model
LLO	Low-Level Orchestrator
LoP	Level of Privacy
LoR	Level of Resilience
LoTw	Level of Trustworthiness
MANO	Management and Orchestration
MEC	Multi-access Edge Computing
Meta-OS	Meta-operating system
MFA	Multi-Factor Authentication
MIMO	Multiple-Input Multiple-Output
ML	Machine Learning
MLOps	Machine Learning Operations
N/A	Not Available
NDT	Network's Digital Twin
NEA	New Encryption Algorithm
NF	Network Function
NFV	Network Functions Virtualization
NIA	New Integrity Algorithm
NICs	Network Interface Cards
NLG	Natural Language Generation
NLP	Natural Language Processing
NRF	Network Repository Function
NSN	Network Service Node
ONAP	Open Network Automation Platform
OpEx	Operational Expenditure
OSC	Open Source Community
OSM	Open-Source Mano
OSS	Operations Support System
PaaS	Platform-as-a-service
PIL	Platform Intelligence
PNFs	Physical Network Functions
POP	Pipeline Orchestration Platform
PS	Privacy Score
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RBAC	Role-Based Access Control
RBO	Regulatory Body Organisation
REST	Representational State Transfer
RNN	Recurrent Neural Network
RTO	Research and Technology Organisation
SBA	Service Based Architecture
SCP	Service Communication Proxy
SDN	Software Defined Network
SDO	Standard Development Organisation
SDP	Software Defined Perimeter

SDPGW	Software Defined Perimeter GateWay
SFC	Service Function Chaining
SHAP	SHapley Additive exPlanations
SLAs	Service Level Agreements
SMF	Session Management Function
SOA	Service-Oriented Architecture
SSI	Self-Sovereign Identity
SSP	Security Service Provider
STO	Secure and Trustable Orchestration
SUCI	Subscription Concealed Identifier
SUPI	Subscriber's Permanent Identifier
TCO	Total cost of ownership
TEEs	Trusted Execution Environments
TF	Trust Function
TLS	Transport Layer Security
TS	Tier-1 Supplier
TV	Telecom Vendor
UCAN	User-Centric Access Network
UCN	User-centric network
UE	User Equipment
UPF	User Plane Function
USN	User Service Node
VC	Verifiable Credential
VM	Virtual Machine
VNF	Virtual Network Function
VR	Virtual Reality
XAI	eXplainable AI
XR	Extended Reality
ZT	Zero Trust

EXECUTIVE SUMMARY

The SAFE-6G project presents a comprehensive, user-centric trustworthiness framework for distributed 6G networks, spanning the edge-cloud continuum.

The most significant paradigm adjustments in the envisioned user-centric 6G system are the shift from a security-only focus to a broader scope of native trustworthiness, clarifying that the term "trustworthiness" refers to a holistic approach, including safety, security, privacy, resilience and reliability. SAFE-6G aims at implementing such a trustworthy framework to support and enhance the trust over the next evolution of the 5G network and the managed resources.

This deliverable establishes the SAFE-6G reference architecture, outlining key components, design principles, and specifications that will inform the development of the rest of the project. Taken as a whole, these elements provide a framework for the project's progress. Establishing the SAFE-6G architecture, outlining crucial elements, and formulating requirements for an end facility that will serve as a guide for numerous tasks are among the main goals. This deliverable outlines how the Service-Based Architecture (SBA) will develop toward a user-centric core, outlines how architectural components will interact, and evaluates the tools that can be used to realize those components.

This document is the second deliverable of WP2 - *D2.2: Overall SAFE-6G Framework and Reference Architecture Design* and after this one follow *D2.3: Metaverse use-cases definition with virtual-assistant for user-centric configuration* (due in M12/December 2024) and *D2.4: 6G Trustworthiness KPI & KVI definition and validation methodology* (due in M14/February 2024).

KEYWORDS

6G, Level of Trustworthiness, Reference Architecture, User Intent, LLM, Cognitive Coordination, User Centric, Trust Functions, Machine Learning

TABLE OF CONTENTS

List of FIGURES

List of TABLES

1 INTRODUCTION

The SAFE-6G project aims at providing an end-to-end cognitive trustworthiness framework for user-centric distributed 6G networks over the edge-cloud continuum.

The main goal of this deliverable is to:

- provide the SAFE-6G reference architecture.
- provide description of the components to be developed.
- define the design and specifications for the end facility (that will serve as a blueprint to guide the development of components in the different tasks).
- define the evolution path of the SBA architecture towards the user-centric core.
- define the required interactions among architectural components.
- study the tools that can be adopted for the realization of the components.

As this description shows it, the produced deliverable is critical for the project, as it will guide the development of all the components, while ensuring coherence with the global reference architecture.

1.1 DELIVERABLE CONTEXT

Item	***Description***
Objectives	**O.1**: Design, build and release a zero-touch holistic E2E cognitive trustworthiness framework for user-centric distributed 6G architectures over the (far) edge-cloud continuum, capable of enabling and supporting the deployment of trusted instances/slices of the user/human-centric 6G system driven by the user's intent and utilizing distributed AI/ML techniques across the entire ecosystem. **O.4**: Follow the cloud-native paradigm over the edge-cloud continuum for the whole design and development of the whole SAFE-6G framework components and the user-centric distributed 5G/6G core network over the edge cloud continuum.
Work plan	D2.2 receives input from: • T2.1: in this task, the consortium members developed a reference blueprint architecture, involving multiple blocks and components interacting with each other. These first ideas have been developed and enriched for this deliverable. D2.2 will drive development of all tasks in: • WP3: User-centric Distributed 6G Core over Edge-Cloud Continuum with MLOps • WP4: AI-driven 6G Trustworthiness Functions and Cognitive Coordination • WP5: Integration, validation, and pilots
Milestones	This deliverable contributes to the realization of *MS4. 6G Trustworthiness requirements and overall design of SAFE-6G framework*, that will be achieved in M10.

Deliverables	In the same spirit as D2.1, deliverable D2.2 will serve as a reference for consortium members, to ensure coherence with SAFE-6G architectural view during the whole duration of the project.

Table 1: Deliverable 2.2 context

1.2 THE RATIONALE BEHIND THE STRUCTURE

This deliverable aims to present the updated SAFE-6G architecture, with regards to the one outlined in the proposal submission, integrating also the components developed under the technical work packages of the project.

In terms of structure, the document begins with the introduction of the methodology adopted by SAFE-6G for the architecture definition, continues with the project specific rationale and then proceeds with the overall architecture presentation. For this, six different architecture perspectives have been considered and are documented. The document further breaks down the essential building blocks and components, the interfaces and communication protocols, illustrating how each contributes to the overall architecture of SAFE-6G concept.

More specifically, the document is structured as follows:

1. Introduction: presents the objectives of the deliverable and introduces the content/structure and expected outcomes.

2. Architecture definition methodology: describes the approach adopted for the definition of the project architecture and the different architecture views that have been considered and the state-of-the-art of the key principles adopted in SAFE-6G.

3. SAFE-6G Rationale briefly explains the holistic approach aimed to design, develop, and validate a 6G-ready native trustworthiness framework.

4. SAFE-6G Reference Architecture: introduces the architecture in terms of the pre-defined different views.

5. SAFE-6G Overall building blocks and Components: explains the key components that comprise the trust framework, such as the cognitive coordinator, the trust functions, the MLOps framework and the Edge cloud continuum.

6. Conclusion: provides a summary of the key points addressed in the deliverable.

2 ARCHITECTURE DEFINITION METHODOLOGY

Mobile networks are complex deployments in which several systems can coexist, at Radio Access Network level (e.g., MIMO, RU/CU, MEC), transport level (e.g., Fiber fronthaul network, Satellite backhaul network, Inter-MNO), core (e.g., SA, NSA) and operational/management levels (e.g., OSS/BSS, MANO, Roaming, MNO applications). SAFE-6G project aims at implementing a trustworthy framework to support and enhance the trust over the next evolution of the 5G network and the managed resources.

Therefore, the SAFE-6G reference architecture should be designed considering interoperability with the rest of the systems (i.e., complying with existing standards of the 5G and beyond ecosystem), aiming at smoothing its future adoption. The methodology for its definition will follow existing standards and best practices that are usually leveraged for describing software systems, serving both as an entry point for external, potential users and adopters of the system as well as a reference document for the technical designs and implementations that will be realized in the project.

A software reference architecture is a generic architecture for a class of system that is used as a foundation for the design of concrete architectures from this class [1] . The SAFE-6G reference architecture will follow the guidelines and concepts of the standard ISO/IEC/IEEE 42010 [2] for software architectures, which addresses the creation, analysis, and sustainment of architectures of systems by using architecture descriptions. The summary diagram of the concepts managed, and their relationships are presented in Figure 1. The main concepts of interest for defining the SAFE-6G architecture are listed here:

- **Entity of interest:** While according to the standard it can be of different types (e.g., solution, system, subsystem, process, application, product line, etc.), it refers to the entity whose architecture is under consideration. In the present case, SAFE-6G system.
- **Stakeholder**: an individual, team, organisation, or classes thereof, having an interest in an *entity of interest*. It includes non-technology stakeholders (such as acquirers and end users) and technology-focused ones (such as developers, system administrators, and maintainers). An exhaustive list of stakeholders related to the SAFE-6G architecture can be found in D2.1-*Definition of Technical Requirements for User-centric 6G Trustworthiness*.
- **Concern**: topic of interest of one or many stakeholders (related to the entity of interest, environment, scenario, situation or use case), which can manifest as a need, goal, risk, expectation, responsibility, requirement, design constraint, assumption, architecture decision, quality attribute, dependency, or other issue. A list of stakeholders' concerns related to SAFE-6G (including potential roles) can be found in D2.1.

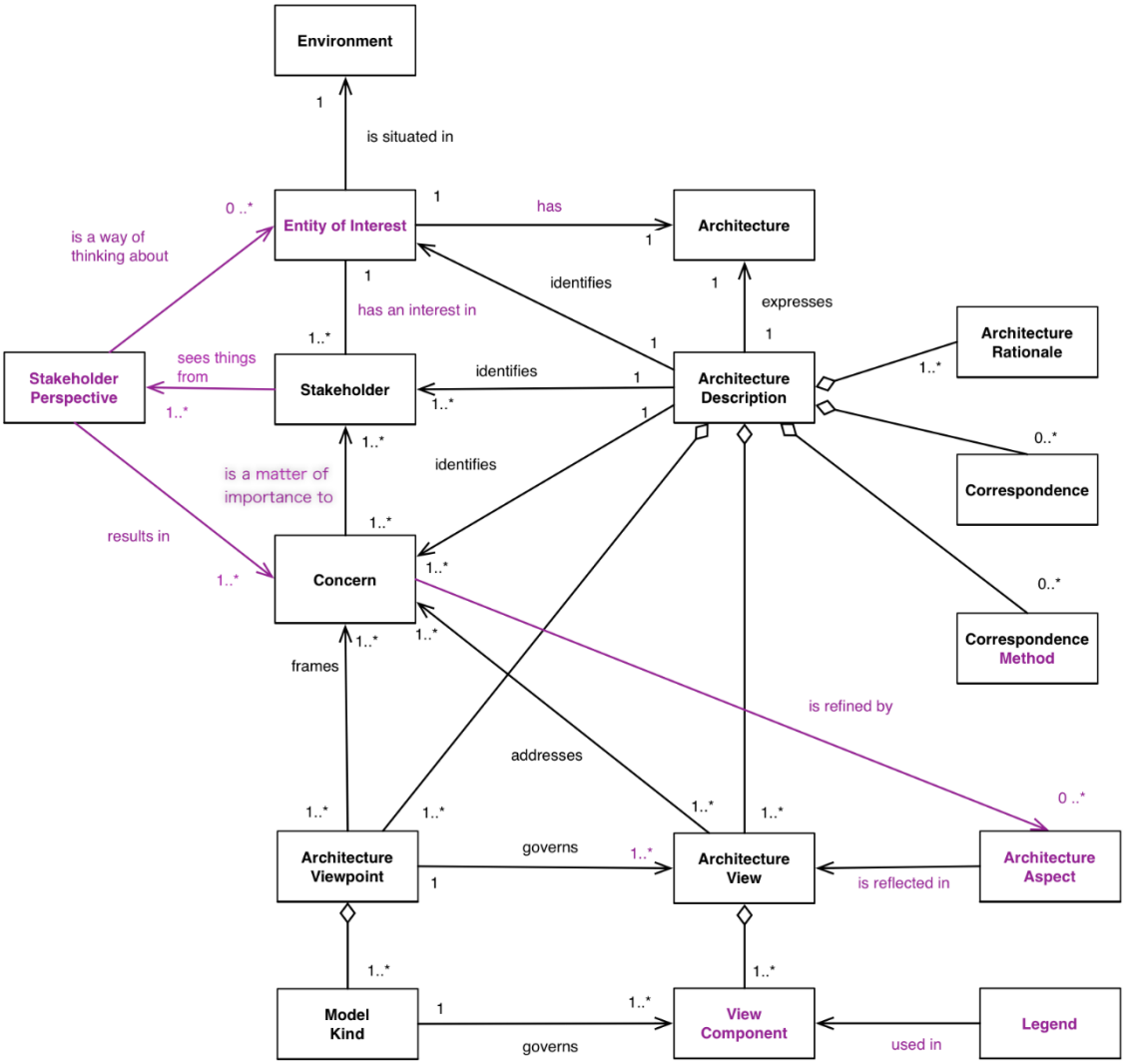

Figure 1: Conceptual model of an architectural description defined in ISO/IEC/IEEE 42010 [2]

- **View**: work product expressing the architecture of an entity of interest from the perspective of specific system concerns.
- **Viewpoint**: set of conventions for the creation, interpretation, and use of an architecture view to frame one or more concerns. Being a common practice for the sake of reducing the complexity of a reference architecture, the terms View and Viewpoint will be used interchangeably, encompassing their expected aspects.
- **Less relevant definitions and linked concepts:** Every Entity of interest inhabits its **Environment**, which acts upon it and vice versa. **Architecture Descriptions** (AD) are comprised of AD Elements. **Correspondences** are used to identify or express named relations within and between AD elements. Creating an Architecture involves making **Architecture Decisions**.

It should be highlighted that the standard does not mandate any specific architecting process, but rather the bases and conceptual guidelines to produce one. With the stakeholders and their concerns already identified in D2.1, SAFE-6G has made the following architectural decision: defining the views

required to represent the reference architecture. The following views, documented afterwards in the deliverable, have been considered:

1. **High-level view**: It describes basic interactions, relationships, and dependencies among the high-level modules that compose the system and its environment.
2. **Functional view**: Cornerstone of most reference architectures, it describes system's functional elements, their responsibilities, interfaces, and primary interactions.
3. **Process view**: It deals with the dynamic aspects of a system, describing system processes and their interactions, and focuses on the run time behaviour of the system.
4. **Data view**: It represents the way that the system manages the data, including their gathering, processing, distribution, storage and presentation.
5. **Deployment view**: Describe the aspects to consider when deploying a system into its target environment.
6. **Business view**: It addresses the business processes, organizational structures, roles, responsibilities, and strategic objectives that the system supports.

Notice that views are not restricted to technical aspects, but can others (e.g., social, economic, etc.). The relationship among the views, inspired on Kruchten's 4+1 view model [3] but adapted to the views leveraged by SAFE-6G, can be seen in the following Figure 2:

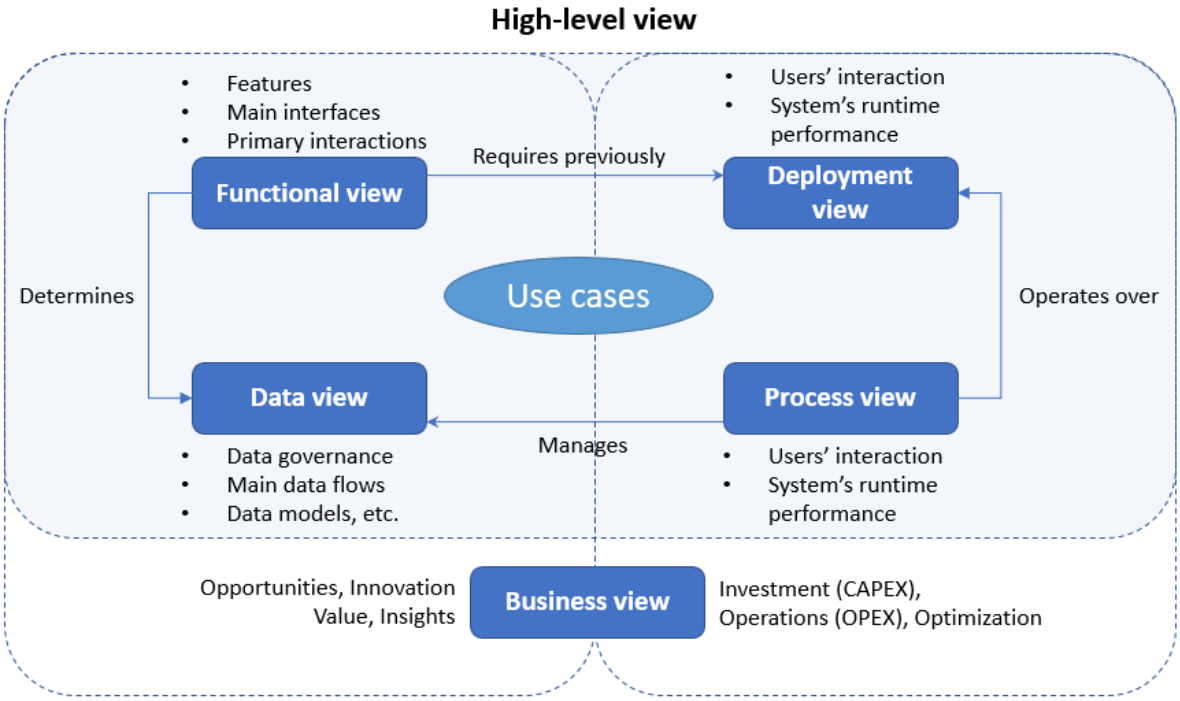

Figure 2: Relation among architecture views

2.1 STATE OF THE ART

2.1.1 STANDARDIZATION LANDSCAPE

2.1.1.1 ISO/IEC 27001

ISO/IEC 27001 [4] is the world's most widely recognized standard for Information Security Management Systems (ISMS). For companies handling sensitive information, ISO 27001 provides a methodological approach to safeguarding sensitive information from misuse, loss, or illegal access. The standard provides recommendations for enterprises of all sizes and industries on how to create, implement, maintain, and continuously improve an information security management system, having adopted strong safeguards and controls to ensure data integrity, confidentiality, and availability. Conformance with ISO/IEC 27001 indicates that an organization or business has implemented a system to manage risks associated with the security of data owned or managed by the enterprise, and that this system adheres to all the best practices and principles. For 5G networks, ISO/IEC 27001 is particularly relevant because it helps organizations establish, implement, maintain, and continually improve their information security management. This is crucial in the 5G context, where vast amounts of data are transmitted and processed, making robust security measures essential to protect against data breaches, cyber-attacks, and other security threats. By adhering to ISO/IEC 27001, organizations can ensure that their 5G infrastructure is secure, reliable, and compliant with international security standards.

ISO/IEC 27001 requires organizations to systematically examine their information security risks, considering threats, vulnerabilities, and impacts. This involves identifying potential security risks and evaluating their likelihood and impact on the organization. Based on the risk assessment, organizations must design and implement a coherent and comprehensive suite of information security controls. These controls can include risk avoidance, risk transfer, risk mitigation, or risk acceptance. The goal is to address risks that are deemed unacceptable and ensure that the organization's information security needs are met on an ongoing basis.

ISO/IEC 27001 includes Annex A, which provides a list of 93 control objectives and controls. These controls cover various aspects of information security, such as access control, cryptography, physical and environmental security, and incident management. Organizations can select and implement controls from Annex A based on their specific risk assessment and security requirements.

In the context of 5G, **ISO/IEC 27001 helps ensure that the vast amounts of data transmitted and processed by 5G networks are protected against unauthorized access, breaches, and other security threats**. This is crucial for maintaining user trust and complying with data protection regulations. The standard's emphasis on risk assessment and treatment is particularly relevant for 5G networks, which face a wide range of security threats. By systematically identifying and addressing these risks, organizations can enhance the security and resilience of their 5G infrastructure. Adhering to ISO/IEC 27001 helps organizations demonstrate their commitment to information security, which is essential for building trust with customers, partners, and regulators. This is especially important in the 5G ecosystem, where security and privacy are critical concerns.

2.1.1.2 3GPP TS 33.501

3GPP TS 33.501 [5] is a technical specification that outlines the security architecture and procedures for 5G systems. It **addresses key security aspects such as authentication, confidentiality, integrity protection, and privacy**. This standard is critical for 5G because it introduces advanced security features designed to protect the network and its users. For example, it includes a unified authentication framework that enhances the security of user access to the network. It also provides secure key management protocols to ensure that encryption keys are handled securely. Additionally, TS 33.501 covers the security of network slicing, which allows multiple virtual networks to operate on a single physical infrastructure, each with its own security requirements. This is essential for supporting diverse 5G use cases, from consumer mobile services to industrial Internet of Things (IoT) applications. TS 33.501 builds on the security foundations laid by previous generations of mobile networks (2G, 3G, 4G) but introduces several enhancements to address the unique challenges and requirements of 5G.

For example, the Unified Authentication Framework involves mutual authentication between the user equipment (UE) and the network (primary authentication). It uses the Authentication and Key Agreement (AKA) protocol, specifically the 5G-AKA or EAP-AKA' (Extensible Authentication Protocol) methods. These protocols ensure that both the UE and the network can verify each other's identities, preventing unauthorized access. For external data networks, the secondary authentication process is foreseen, which provides an additional layer of security by requiring the UE to authenticate with an external authentication server, typically using protocols like EAP-TLS (Transport Layer Security).

TS 33.501 specifies the use of advanced encryption algorithms such as 128-NEA (New Encryption Algorithm) and 256-NEA. These algorithms protect the confidentiality of user data as it travels over the air interface.

In terms of integrity protection, TS 33.501 ensures that the data has not been tampered with during transmission. Algorithms like 128-NIA (New Integrity Algorithm) and 256-NIA are used to provide integrity protection for both control and user plane data.

In terms of user privacy, TS 33.501 introduces the concept of Subscription Concealed Identifier (SUCI), which is a concealed version of the subscriber's permanent identifier (SUPI). The SUCI is generated using a public key of the home network, ensuring that the SUPI is not exposed over the air interface. Similarly, the standard also includes mechanisms to protect the location information of users, preventing unauthorized tracking and location-based attacks.

2.1.1.3 ETSI TS 103 305

ETSI TS 103 305 provides guidelines for implementing security in network functions virtualization (NFV) environments, which are a key component of 5G networks. NFV allows network services to be virtualized and run on standard hardware, rather than proprietary hardware. This flexibility is crucial for the scalability and efficiency of 5G networks. However, it also introduces new security challenges, as virtualized environments can be more vulnerable to cyber threats. ETSI TS 103 305 addresses these challenges by providing recommendations for secure deployment, operation, and management of NFV environments. It ensures that virtualized network functions are resilient against attacks and that

security is maintained across the entire lifecycle of the NFV infrastructure. This standard is vital for maintaining the trustworthiness of 5G networks, as it helps protect against threats that could compromise the integrity and availability of network services.

ETSI TS 103 305 outlines specific technical measures to detect and prevent cyber-attacks. These measures include intrusion detection systems (IDS), intrusion prevention systems (IPS), and advanced threat detection techniques. By considering these controls, 5G networks can identify and mitigate potential threats before they cause significant damage. The standard also provides guidelines for responding to and mitigating the impact of cyber-attacks. This includes incident response plans, disaster recovery procedures, and business continuity strategies, which could assist a 5G system to recover from a security incident and maintain service availability.

ETSI TS 103 305 emphasizes the importance of encrypting data both in transit and at rest. This ensures that sensitive information is protected from unauthorized access and interception. The standard provides guidelines for implementing robust access control mechanisms. This includes role-based access control (RBAC), multi-factor authentication (MFA), and least privilege principles.

ETSI TS 103 305 addresses the unique security challenges associated with virtualized environments. This includes **securing the hypervisor, virtual machines (VMs), and virtual network functions (VNFs)**. Thus, the standard outlines best practices for securely deploying and managing NFV environments. This includes guidelines for secure configuration, patch management, and monitoring. Since virtualization is a critical enabler of the scalability and flexibility of 5G networks, ETSI TS 103 305 ensures that these virtualized functions are secure, allowing 5G networks to scale without compromising security. The standard's focus on detection, prevention, response, and mitigation helps ensure that 5G networks are resilient against cyber threats. This is essential for maintaining the reliability and availability of 5G services, especially for critical applications such as industrial IoT and smart cities.

2.1.1.4 ITU-T Y.3060

Y.3060 (Autonomous networks — Overview on trust) provides a structured foundation for thinking about trust in autonomous and AI-driven networks, explaining why trust is essential as networks gain autonomy and intelligence. The Recommendation was published as part of the ITU-T Y.3000 series and sets out the background, core concepts and the high-level scope for trust work in network autonomy.

The Recommendation **defines trust concepts** specifically for autonomous networks, distinguishing between trust attributes, trust provisioning and the lifecycle of trust in operational networks. It lays out **basic principles** for trusted autonomous networks, such as transparency, accountability, verifiability and adaptability, and explains how these principles apply across network functions and management planes.

A central contribution of Y.3060 is an **overall workflow model** for trusted autonomous networks: how trust requirements are captured, how trust policies are translated into operational controls, how telemetry and evidence are collected, and how verification/validation loops close the trust lifecycle.

This model is intended to guide architects and operators in turning high-level trust objectives into implementable controls and monitoring processes.

Y.3060 includes **illustrative use cases** showing how trust considerations arise in real deployments (for example, autonomous orchestration, cross-domain service composition and AI-assisted control loops). These use cases help stakeholders map abstract trust principles to concrete scenarios and identify where additional standards, testing or governance are needed.

The Recommendation builds on prior ITU work and outputs from the Focus Group on Autonomous Networks; it was prepared by a multi-author team and synthesizes academic, industry and standards perspectives on trust in autonomous systems. Y.3060 is positioned to interoperate with other ITU and standards outputs that address trust provisioning, security and autonomous network use cases.

Y.3060 is **practical rather than prescriptive**: it does not mandate specific protocols but provides a conceptual and procedural framework that operators, vendors and standards bodies can use to design trustable autonomy. Its scope explicitly covers the introduction of trust concepts into network intelligence, the principles to follow, and the development steps for trusted autonomous networks—making it a useful reference for architects and policy teams planning AI-driven network features.

2.1.2 Regulatory Landscape

2.1.2.1 NIS2

The **NIS2 Directive** (Directive (EU) 2022/2555)[6] is a European Union (EU) directive aimed at enhancing the cybersecurity of network and information systems across the EU. It replaces the original NIS Directive (Directive (EU) 2016/1148) and introduces more stringent requirements to address the evolving cyber threat landscape. NIS2 aims to improve the overall level of cybersecurity in the EU by expanding the scope of sectors covered, enhancing incident reporting obligations, and strengthening security requirements.

NIS2 expands the scope of the original directive to include more sectors and services that are critical to the economy and society. This includes sectors such as energy, transport, banking, health, digital infrastructure, and public administration. By covering a wider range of sectors, NIS2 ensures that **more entities are required to implement robust cybersecurity measures**. The directive distinguishes between essential and important entities, with essential entities being subject to stricter security requirements and oversight. This classification helps prioritize resources and efforts towards protecting the most critical infrastructure.

NIS2 mandates that organizations implement comprehensive risk management practices. This includes identifying and assessing cybersecurity risks, implementing appropriate security measures, and ensuring continuous monitoring and improvement. Organizations are also required to **establish governance frameworks to oversee cybersecurity efforts and ensure accountability**.

NIS2 places a strong emphasis on securing the supply chain. Organizations are required to **assess and manage risks associated with third-party vendors and service providers**. This helps prevent vulnerabilities from being introduced through external partners.

NIS2 expanded scope and stringent security requirements help ensure that 5G networks are protected against cyber threats, enhancing their reliability and trustworthiness. The emphasis on supply chain security is particularly relevant for 5G, where components and services are often sourced from multiple vendors. By requiring organizations to assess and manage supply chain risks, NIS2 helps prevent vulnerabilities that could compromise the security of 5G networks. The enhanced incident reporting and coordination mechanisms established by NIS2 are crucial for 5G networks, which need to respond quickly to cyber threats to maintain service continuity.

2.1.2.2 EU 5G CYBERSECURITY TOOLBOX

The **EU 5G Cybersecurity Toolbox** [7] is a comprehensive set of measures developed by the EU to ensure the secure deployment and operation of 5G networks across member states. It was created in response to the EU coordinated risk assessment of 5G network security, aiming to mitigate identified risks and enhance the overall cybersecurity posture of 5G infrastructures. The toolbox provides both strategic and technical measures to address various cybersecurity challenges associated with 5G and recommends enhancing the regulatory powers of national authorities to scrutinize network procurement and deployment. This includes the ability to impose restrictions on high-risk suppliers and ensure that critical network components are sourced from trusted vendors.

The toolbox addresses risks related to non-technical vulnerabilities, such as dependencies on specific suppliers or geopolitical factors and suggests promoting a diverse and sustainable 5G supply chain to avoid long-term dependency risks.

On the technical domain, the toolbox outlines specific security requirements for network operators, including the implementation of robust encryption, secure authentication mechanisms, and regular security audits, which help protect the integrity and confidentiality of data transmitted over 5G networks. Moreover, it provides guidelines for securing the 5G supply chain, including conducting thorough risk assessments of suppliers and implementing stringent security requirements for third-party components and services. Similarly, the toolbox includes measures for enhancing incident response and recovery capabilities. This involves establishing incident response teams, developing incident response plans, and conducting regular exercises to test and improve response procedures.

In this context, the toolbox encourages investment in research and development to advance cybersecurity technologies and solutions. This includes **supporting innovation in areas such as artificial intelligence (AI) and machine learning (ML) to enhance the security of 5G networks**, providing a holistic approach to 5G security, addressing both strategic and technical aspects. This comprehensive framework helps ensure that 5G networks are protected against a wide range of cyber threats.

2.1.2.3 ENISA 5G TOOLKIT

The ENISA 5G Toolkit [8] is a comprehensive set of guidelines and best practices developed by the European Union Agency for Cybersecurity (ENISA) to enhance the cybersecurity of 5G networks. This toolkit is part of the broader EU 5G Cybersecurity Toolbox, previously described, which was established to address the security challenges associated with the deployment of 5G networks across

the EU. The toolkit provides strategic, technical, and supporting measures to mitigate cybersecurity risks and ensure the resilience of 5G infrastructures.

The ENISA 5G Toolkit emphasizes the importance of conducting thorough risk assessments to identify potential threats and vulnerabilities in 5G networks. The toolkit recommends implementing stringent security requirements for suppliers and conducting regular audits to verify compliance. This helps prevent the introduction of vulnerabilities through third-party components and services.

Towards this, the ENISA 5G Toolkit includes a comprehensive Security Controls Matrix, which provides a detailed list of security controls and best practices for 5G networks. This matrix covers various aspects of network security, including access control, encryption, and incident response. It serves as a practical tool for organizations to implement and maintain robust security measures. The toolkit also addresses the security challenges associated with NFV, a key technology in 5G networks, by providing guidelines for securing virtualized network functions, including recommendations for secure configuration, monitoring, and management of NFV environments.

The ENISA 5G Toolkit provides a holistic approach to 5G security, addressing both strategic and technical aspects. This comprehensive framework helps organizations implement robust security measures that can protect 5G networks from a wide range of cyber threats. By promoting risk assessment, supply chain security, and collaboration, the toolkit enhances the overall resilience of 5G networks. This is crucial for maintaining the reliability and availability of 5G services, especially for critical applications such as smart cities and industrial IoT.

2.1.3 Multidimensional Trustworthiness and AI Governance Landscape

While the foundational security architectures provided by 3GPP TS 33.501 and ISO/IEC 27001 are essential for the protection of confidentiality and integrity, the evolution toward 6G cyber-physical systems necessitates a broader conceptualization of network dependability that transcends traditional cybersecurity. The standardization landscape has increasingly recognized that "trustworthiness" is not a singular attribute but a composite property, as formalized in **ISO/IEC TS 5723:2022** [9] , which defines trustworthiness as the verifiable ability to meet stakeholder expectations through a convergence of safety, security, privacy, reliability, and resilience. This holistic perspective is echoed in the **NIST Framework for Cyber-Physical Systems** [10] , which categorizes these attributes as critical top-level concerns that must be balanced against one another, acknowledging that optimizing for one dimension, such as high-overhead encryption for security, may inadvertently degrade another, such as latency-sensitive reliability. This multidimensional approach validates the SAFE-6G rationale of shifting from a security-only focus to a native trustworthiness framework, where the interplay between these diverse attributes must be managed dynamically rather than statically.

In the domain of Artificial Intelligence, which SAFE-6G leverages through its Cognitive Coordinator and AI agents, the regulatory and standardization landscape is shifting rapidly from theoretical ethics to operational governance. **ISO/IEC 42001:2023** [11] has emerged as a dedicated international standard for AI management systems, mandating that organizations implement continuous monitoring, risk assessment, and transparency mechanisms throughout the AI lifecycle to address specific

vulnerabilities such as bias, model drift, and lack of explainability. Concurrently, the **EU AI Act** [12] establishes a risk-based legal framework that imposes strict obligations on high-risk AI systems, which likely include critical 6G network control functions, requiring robust data governance, human oversight, and detailed documentation to ensure traceability. These developments underscore the importance for SAFE-6G to incorporate eXplainable AI (XAI) and MLOps not merely as technical enhancements, but as essential compliance mechanisms that align network automation with emerging requirements for transparent and accountable algorithmic decision-making.

Furthermore, the specific dimensions of resilience and reliability are being standardized through frameworks that emphasize automated recovery and service continuity in distributed environments. **ETSI GS ZSM (Zero-touch Network and Service Management)** [13] represents the state of the art in network automation, defining reference architectures for closed-loop management that enable systems to detect anomalies and autonomously trigger remediation actions, thereby ensuring resilience without human intervention. This aligns with the SAFE-6G objective of utilizing distinct Trust Functions to maintain service levels, moving beyond simple redundancy toward active, AI-driven self-healing. Parallel to this, **3GPP's work on Ultra-Reliable Low Latency Communications (URLLC)** continues to refine the metrics for reliability and availability, pushing toward deterministic networking capabilities that are requisite for the industrial metaverse, and mission-critical applications envisioned in the SAFE-6G use cases.

Finally, the dimensions of safety and privacy require specialized frameworks that address the physical and data-centric risks inherent in a user-centric 6G continuum. **IEC 61508** [14] regarding functional safety is increasingly relevant as 6G networks integrate with physical control systems, necessitating that network architectures possess fail-safe mechanisms to prevent cyber-attacks from escalating into physical hazards, a principle integrated into the SAFE-6G Safety Trust Function. Regarding privacy, **ISO/IEC 27701** [15] extends general information security standards to specifically govern Privacy Information Management Systems (PIMS), providing the guidelines for managing Personally Identifiable Information (PII) in compliance with regulations like GDPR. By integrating these diverse standardization streams, the SAFE-6G framework addresses the current gap in the state of the art, where these trustworthiness pillars are typically managed in isolation, offering instead a unified, intent-based coordination layer that dynamically optimizes the trade-offs between safety, security, privacy, resilience, and reliability.

2.1.4 Technology Domains Landscape

This section provides an overview of the current state-of-the-art technologies and methodologies in the fields Cloud Native, Orchestration, Architecture, MLOps and Chatbot domains that are crucial to SAFE-6G User-Centric architecture.

2.1.4.1 Cloud Native Domain

The industry is moving from traditional Physical Network Functions (PNFs) and Virtual Network Functions (VNFs) to Cloud Native Network Functions (CNFs). This transition is driven by the need for greater scalability, flexibility, and efficiency in network operations. Key principles such as microservices architecture, containerization, and orchestration are being adopted. These principles enable more agile and resilient network services.

Organizations like ETSI and 3GPP are working on standards to ensure interoperability and consistency in the deployment of Cloud Native technologies across different platforms and vendors. Open-source communities, including ONAP and CNTT, are playing a significant role in developing and promoting Cloud Native solutions. These contributions help in accelerating innovation and adopting Cloud Native technologies.

The Cloud Native layer has evolved significantly, with its reference implementation initially developed by Google and now advanced by the Cloud Native Computing Foundation (CNCF). CNCF, part of the Linux Foundation, brings together some of the largest open-source projects in the Cloud Native space. Projects such as Kubernetes, Prometheus, Fuentd and Helm. Containers, service meshes, microservices, immutable infrastructure, and declarative APIs enable the creation of loosely coupled systems that are resilient, manageable, and observable, reflecting the cutting-edge practices in the Cloud Native domain.

XGVela's contribution to Cloud Native [16] focuses on platform sharable capabilities, such as common microservices' functional components (e.g., databases, load balancers, firewalls, common NF microservices). These components can be shared among different vendors, applications, and network functions. XGVela includes observability to ensure that the system is transparent, and their performance can be monitored effectively, Platform-as-a-service (PaaS), and Infrastructure-as-a-service (IaaS). XGVeal's works is relatively recent and aims to address the comprehensive nature of being Cloud Native, which encompasses not only containerization but also microservices and DevOps. This brings challenges to the telecom industry in terms of network functions design, delivery, operations, and procurement models. However, XGVela strives to add Cloud Native value to the Telco cloud platform, making it easier to run, create, and manage network functions and applications.

From Service-Oriented Architecture (SOA) for Distribution Grids [17] paper presents a new philosophy for the digitalization and automation of distribution grids, based on a modular architecture of microservices implemented via container technology. This architecture enables a service-oriented deployment of the intelligence needed in the Distribution Management Systems (DMS), moving beyond the traditional view of monolithic software installations in the control rooms. DMS includes functionalities such as state estimation, fault location, isolation, service restoration (FLISR), and forecasting. These applications are often deployed as microservices, enhancing their scalability and flexibility. The paper declares that microservices is a variant of SOA that structures an application as a collection of loosely coupled services and as benefits, an architecture based on microservices enhances modularity, making easier to develop, test, deploy, and scale applications. It also improves fault isolation and enables continuous delivery and deployment.

Building on these principles, SAFE-6G architecture leverages modern concepts and methodologies to ensure a robust, scalable, and efficient 6G network infrastructure. This includes DevOps practices to enable rapid deployment cycles and efficient collaboration between teams, the known concept Cloud Native is defined in the SAFE-6G architecture using containerization and orchestration technologies as Kubernetes to enhances flexibility, scalability, and resource efficiency.

2.1.4.2 ORCHESTRATION DOMAIN

The **Open Network Automation Platform (ONAP)** [18] represents a cutting-edge comprehensive open-source solution for the orchestration, management, and automation of networks. Designed for network operators, cloud providers, and enterprises, ONAP enables the rapid and dynamic instantiation of network elements and services. ONAP's information model and framework utilities are continuously evolving to harmonize with the work of various standard development organizations (SDOs) such as ETSI NFV MANO, TM Forum SID, ONF Core, OASIS TOSCA, IETF MEF, and 3GPP. Integration ONAP into a platform that includes Business Support Systems (BSS), Operations Support System (OSS), end-to-end (E2E) and domain orchestration, while ensuring compliance with mobile networks standards is complex. ONAP operates at different levels and must be integrated into an overall OSS System composed of multiples entities managing the network.

The ESOUPN [19] presents the principles to automate and optimize the orchestration of network functions leveraging cloud-native technologies and Software Defined Network (SDN). The paper outlines the use of cloud-native technologies based on containerization and Kubernetes, facilitating the automated deployment, scaling, and management of network functions. This approach enhances the flexibility and scalability of the network, allowing for efficient orchestration of services. The document explains how the architecture would provide SMF and UPF deployment automatically through external technologies such as OpenShift or Kubernetes, integrated into ETSI Management and Orchestration (MANO) frameworks, Open-Source Mano (OSM), or ONAP. However, they address the challenges of interoperability, scalability, and security and privacy. Regarding interoperability, the organization must be capable of managing heterogeneous environments effectively. For scalability, the orchestration framework must scale efficiently to handle the increased load without compromising performance. The security and privacy challenge requires the orchestration to preserve security through E2E measures, meaning it must implement robust security measures and ensure compliance with privacy and regulatory requirements. Finally, the Key Future directions in the article include enhanced AI integrations and ML into orchestration systems to improve predictive capabilities and enable more sophisticated automation. Also, availability of edge orchestration, where orchestration techniques for edge computing must offer low latency in applications and networks.

As described in the D2.1 SAFE-6G document definition in section 2.5.2, the Cognitive Coordinator involves the orchestration of the SAFE-6G framework to calculate the level of trustworthiness and orchestrate the User-Centric SAFE-6G functions to adapt the 6G core system into a trustworthy system for the user.

2.1.4.3 6G PLANES DOMAIN

The **5G Architectural Design Patterns** [20] paper emphasizes the importance of network slicing, which allows the creation of multiple virtual networks on a shared physical infrastructure. This is crucial for supporting diverse use cases and ensuring efficient resource utilization. Also, the authors discuss advanced RAN design patters that integrate cloud computing, SDN, and NFV which are flexible technologies for creating scalable network architectures. The paper underlines the need for integration various software and hardware technologies to enhance the performance and capabilities of 5G systems. This includes the use of microservices, containers, and orchestration tools. Additionally,

the study underscores the importance of standardization efforts by organizations such as ETSI NFV MANO, TM Forum SID, ONF Core, OASIS TOSCA, IEFT MEF and 3GPP to ensure interoperability and consistency across different platforms and vendors.

This paper [21] focuses on creating a highly autonomous and efficient network management system, where the models define how user intents are captured, interpreted, and translated into network policies, and the network understands and executes user requirements accurately. The Lifecycle Manager manages the process of creation, validation, monitoring, deployment, and refinement of intents, ensuring that the network adapts to changing requirements and conditions. Additionally, they incorporate context-aware mechanisms, allowing decisions to be made in the network based on real-time and environmental conditions, enhancing the efficiency of the network.

On the other hand, the paper integrates NFV and SDN, which provide the flexibility and programmability needed to dynamically manage network resources based on user intents. Furthermore, Service Level Agreements (SLAs) are integrated into the architecture framework to guarantee that the network meets predefined performance and reliability standards. This integration helps in maintaining the quality of service as per user expectations. The paper uses the ESTI ZMS architecture as a reference to align with its architecture shown in the next figure. However, they found challenges in standardization, where there is a need for standardized models and protocols to ensure interoperability between different network components and vendors.

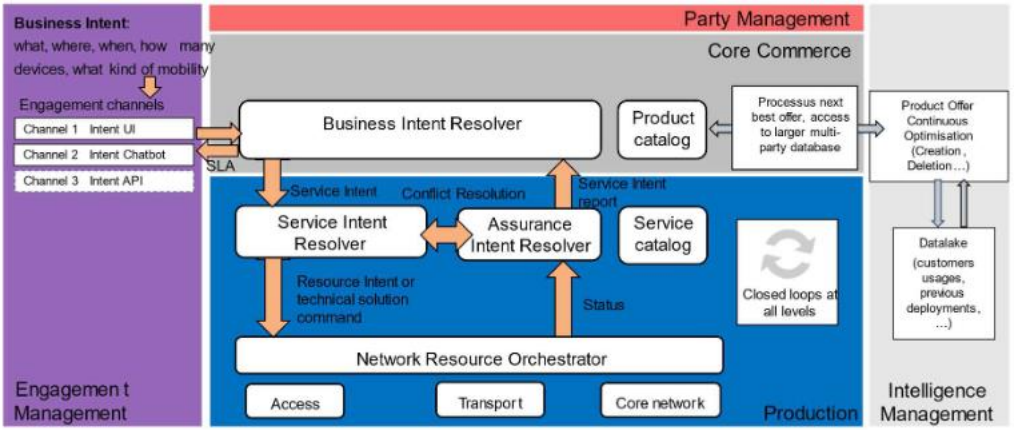

Figure 3: TMF ODA architecture [21] .

In the SAFE-6G architecture, it is important to design and create an architecture methodology that supports all challenges associated with a distributed user-centric architecture. In this context AI plays a crucial role in decision-making within the 6G Core system, ensuring the Level of Trustworthiness (LoTw) in each domain.

2.1.4.4 MLOps Domain

From [22] 5G networks are substantially more complicated than previous generations, requiring advanced technology such as AI to ensure reliable operation. The paper specifies ML and deep learning as essential AI techniques for 5G, helping in tasks such as wireless signal processing, channel

modelling, and resource management. AI can increase network performance, manage resources more efficiently, and improve overall network management and monitoring.

The ML layer uses supervised learning for tasks like traffic prediction and user behaviour analysis. Unsupervised learning helps in anomaly detection and clustering of network data, while reinforcement learning optimizes resource allocation and network slicing by learning from the environment. On the other hand, the Deep Learning layer uses Convolutional Neural Networks (CNNs) for image and video data processing to enhance multimedia services. Additionally, Deep Learning leverages the Recurrent Neural Network (RNN) concept for sequential data analysis, such as predicting user mobility patterns. Finally, it uses Generative Adversarial Networks (GANs) to generate synthetic data, improving training models and enhancing security features.

In conclusion, this paper presents a novel approach to integrating AI and 5G networks to enhance resource consumption efficiency. By leveraging network slicing, it provides network owners with greater control over usage. Additionally, the use of ML improves performance, functionality, and security management. ML and AI can identify patterns related to Quality of Service (QoS) and mobility, predicting congestion in specific locations and zones throughout the day.

A proof of concept of an AI-assisted user-centric 6G network [23] suggests that the network gathers extensive data from user devices, network nodes, and applications in order to be analysed and understand user behaviour, preferences, and requirements. AI and ML models predict user intents and optimize network resources accordingly. The network slicing method creates virtual networks tailored to specific user intents; therefore, each slice can be optimized for different QoS parameters, such as latency, bandwidth, and reliability. Additionally, edge computing deploys resources closer to the user (at the edge of the network), which reduces latency and improves the responsiveness of applications which is crucial for real-time applications like augmented reality (AR) and virtual reality (VR).

As underline in D2.1 SAFE-6G document, MLOps domain plays a key role in the SAFE-6G framework. It ensures that AI models remain in production, and the User-Centric functions continuously train their models. It includes providing methods, interfaces, libraries, models and more, which are all part of SAFE-6G framework.

2.1.4.5 CHATBOT DOMAIN

In the [24] paper, they propose a system architecture that reacts to customer intents. Various standardization bodies, such as TM Forum and ETSI, define intents as high-level expressions of user or system expectations. These intents focus on “what” the system should achieve rather than “how” to accomplish it. This method enables systems to autonomously determine the best way to fulfil the specified intents.

To interact with customers/users, as shown in Figure 4, they developed a chatbot that is integrated into the Web UI and designed for non-technical users. It helps define customer intents interactively and intuitively. The chatbot uses Natural Language Processing (NLP) and transformer technology to understand and process customer requests. This includes leveraging models like BERT for improved performance.

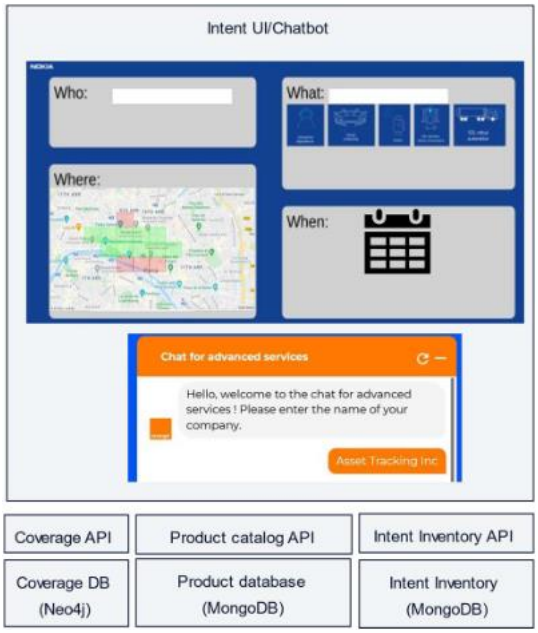

Figure 4: Intent UI/Chatbot [24]

Additionally, the chatbot allows users to describe their use cases in natural language, making it easier to match their needs. The chatbot can recognize and classify product-related entities even when users do not provide exact product names, enhancing accuracy and user experience.

It is noted that under the User-Centric field [25] , the importance of designing 6G networks with a focus on the end-user experience ensures that the network meets the diverse needs of users, where intent-based mechanisms play a role in adapting the network based on users' intents and requirements, providing a more personalized and efficient service.

As outlined in the D2.1 SAFE-6G document definition in section 2.5.1, the chatbot is a crucial domain of the SAFE-6G framework. It is the first element in the framework that contributes to enhancing and ensuring trustworthiness in the User-Centric Distributed 6G Core.

2.1.5 Research Initiatives Landscape

2.1.5.1 CONFIDENTIAL-6G

CONFIDENTIAL6G [26] is a phase 1 SNS project, which will develop tools, libraries, and blueprints to ensure confidentiality in 6G. This includes cryptographic enablers as the foundation for building advanced software components, platforms, and applications that enhance secure communication and computing. This will involve using techniques such as secure multi-party computation and federated AI/ML orchestration. The development of future systems will also rely on advanced cryptographic protocols that are resistant to quantum computing attacks, and formal security proofs to ensure the highest level of security.

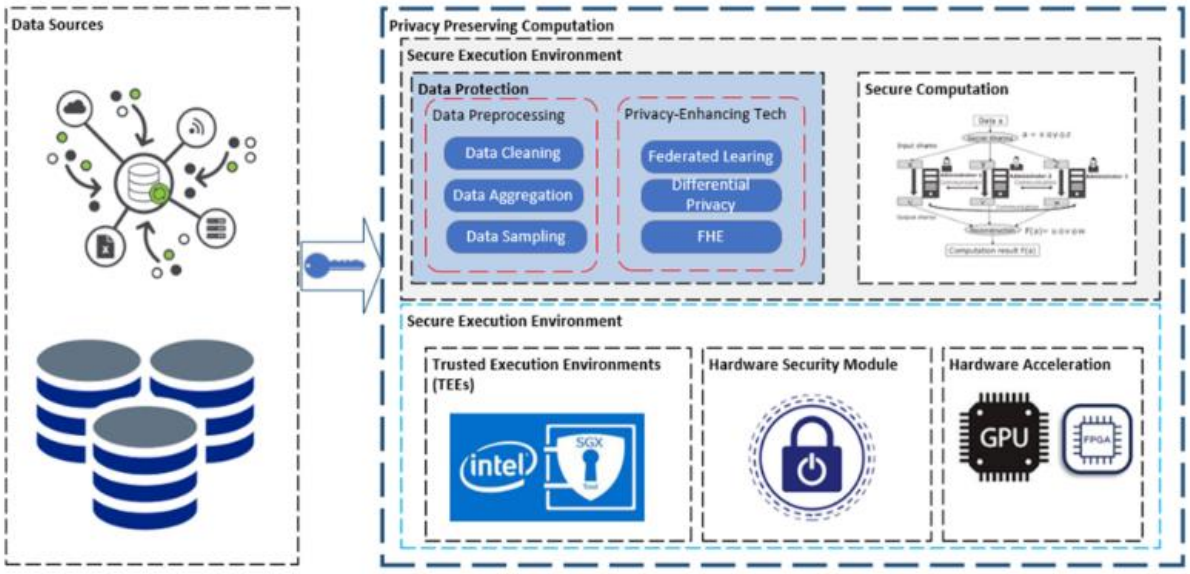

Figure 5: CONFIDENTIAL6G architecture [26]

CONFIDENTIAL6G (Figure 5) emphasizes privacy preservation and security of sensitive data by focusing on protection of data:

- **In use**. This is an unsolved issue with solutions just emerging with Confidential Computing.
- **In transit**. CONFIDENTIAL6G will enhance communication protection with post-quantum cryptography, blockchain technologies and secure data access control and traceability platforms.
- **At the Edge**. CONFIDENTIAL6G will work on specifying the post-quantum cryptographic approach most appropriate to cater for constrained Edge and IoT devices. Additionally, to avoid data movement from the Edge and increase trust and security, Federated AI/ML will be researched.

CONFIDENTIAL6G will base its research on three pillars: Post-quantum cryptography, Confidential Computing and Confidential Communication. CONFIDENTIAL6G will test and validate the developed solutions in three use cases: 1) Predictive maintenance for airline consortium using blockchain-based data sharing platform and federated AI/ML orchestration, 2) Privacy-preserving confidential computing platform that enables mitigation of internal threats for telecom cloud providers and 3) Intelligent connected vehicle, mission-critical services, OTA updates, FL/ML, and vehicle to infrastructure communication.

2.1.5.2 DESIRE-6G

DESIRE6G [27] is a phase 1 SNS project, which will design and develop a novel zero-touch control, management, and orchestration platform, with native integration of AI, to support eXtreme URLLC application requirements. DESIRE6G will re-architect mobile networks through a) its intent-based control and end-to-end orchestration that targets to achieve near real-time autonomic networking; and b) a cloud-native unified programmable data plane layer supporting multi-tenancy. A generic hardware abstraction layer will support the latter for heterogeneous systems. The flexible

composition of modular micro-services for slice-specific implementations and flexible function placement depending on HW requirements will enable granular use case instantiation and service level assurance with minimum resource consumption and maximum energy efficiency.

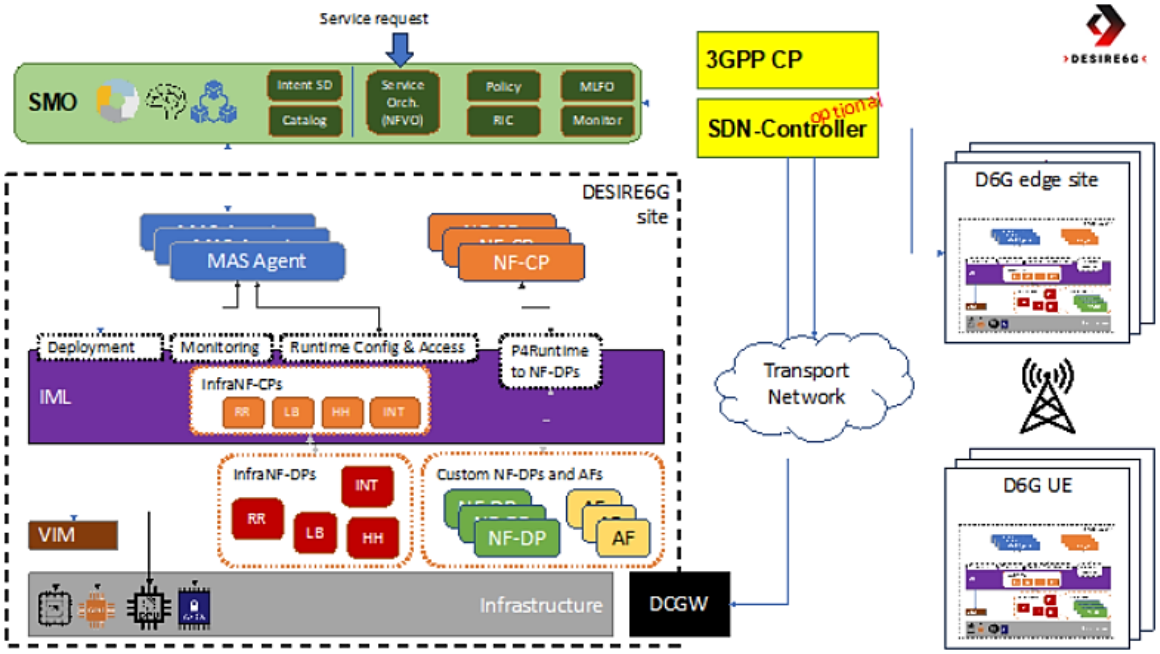

Figure 6: DESIRE6G architecture [27]

The DESIRE6G data, control, management, and orchestration plane is supported by a pervasive monitoring system, extending from the network to the user equipment or IoT terminal (Figure 6). DESIRE6G will employ distributed ledger technology to support a) dynamic federation for services across multiple administrative domains and b) infrastructure-agnostic software security. Finally, DESIRE6G will enable communication-, and energy-efficient distributed AI, at the network edge, while considering application-level requirements and resource constraints. The proposed innovations will be validated by employing a VR/AR/MR and a Digital Twin (DT) application at two different experimental sites.

2.1.5.3 HORSE

The HORSE [28] is a phase 1 SNS project, which envisions the development of a novel platform to address the evolving demands of 6G networks, characterized by advanced softwarisation, high-speed (Gb/s), and sub-THz communications. As 6G embraces new paradigms such as network disaggregation, virtualisation, and multi-vendor ecosystems, HORSE seeks to build a human-centric, open-source, green, and sustainable framework. It will facilitate seamless integration across multiple domains, including proactive threat detection, programmable networking, AI-driven management, semantic communications, and more.

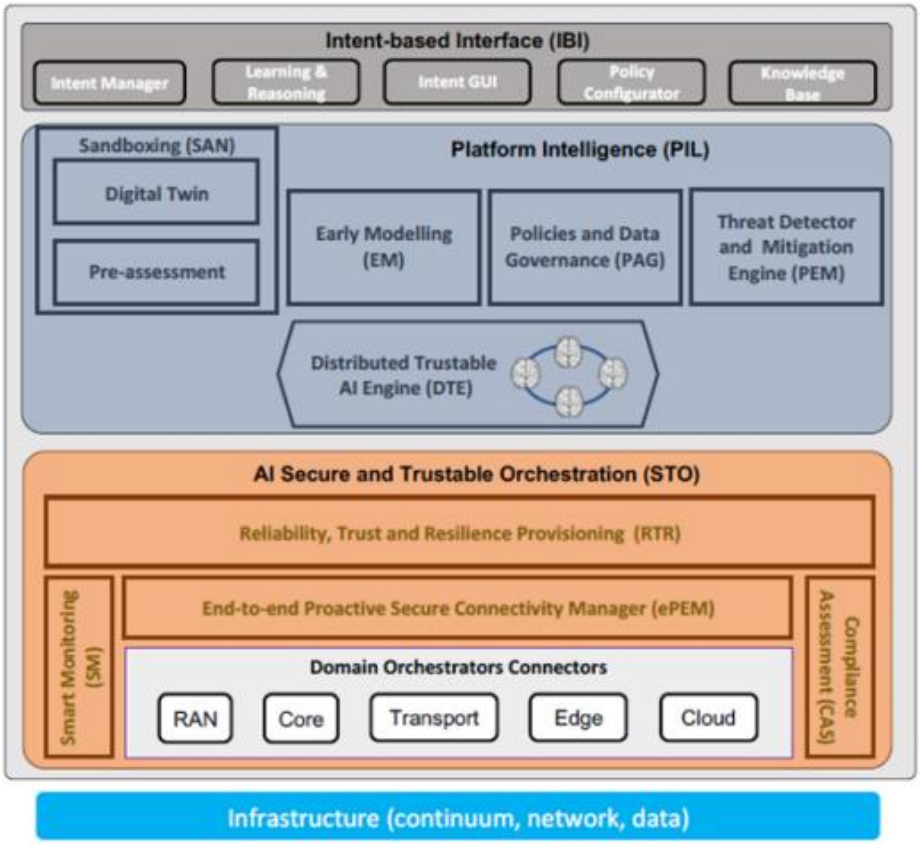

Figure 7: HORSE architecture [28]

Key innovations of HORSE (Figure 7) revolve around designing a multi-layered architecture with the following components:

1. **Intent-based Interface (IBI):** Simplifies network configuration and operations by translating high-level intents from network managers into actionable policies.
2. **Platform Intelligence (PIL):** Incorporates AI-driven modules for predictive security, anomaly detection, and mitigation in real-time as well as in sandboxed, emulated environments.
3. **AI Secure and Trustable Orchestration (STO):** Ensures secure and trustworthy operation, with advanced monitoring and compliance mechanisms.

HORSE's approach leverages digital twins to emulate real-worlds scenarios, predictive analytics for threat detections, and AI models to ensure proactive, intelligent network management. The project also emphasizes security, trust, and privacy by adhering to various industry standards and regulatory frameworks, with a vision to achieve an omnipresent, smart, and resilient network in the future interconnected landscape.

2.1.5.4 PRIVATEER

The PRIVATEER [29] is a phase 1 SNS project, which aims to lead the development of privacy-centric security mechanisms for 6G networks. In the multi-actor environment of 6G, where privacy is crucial for both users and stakeholders, PRIVATEER focuses on ensuring that privacy is preserved throughout the network stack, particularly in security enablers.

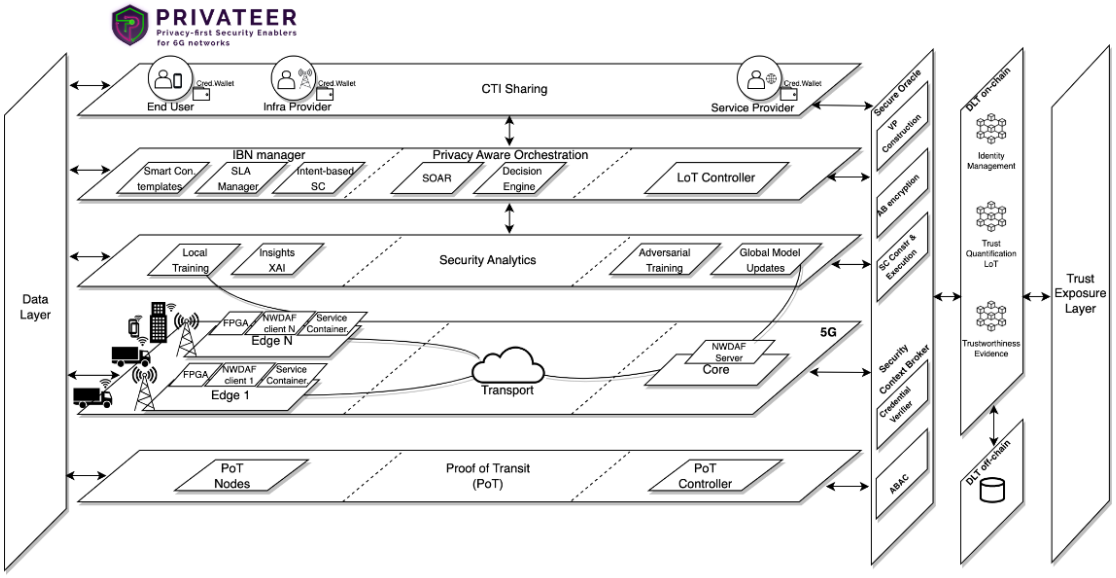

Figure 8: PRIVATEER architecture [29]

Key goals of PRIVATEER (Figure 8) include:

1. **Decentralized Security Analytics:** Moving away from centralized data collection, PRIVATEER promotes robust decentralized security analytics, ensuring that AI mechanisms are resilient to privacy breaches, and data centralization is avoided.
2. **Privacy-Aware Network Slicing and Orchestration:** PRIVATEER introduces the concept of "privacy intent," making privacy a core factor in the lifecycle management of network services. This includes enabling orchestration with privacy considerations in mind.
3. **Distributed Attestation and Identity Checks:** Authentication and integrity verification are performed in a manner that prioritizes privacy, utilizing privacy-friendly mechanisms that limit exposure of sensitive information.
4. **Searchable Encryption for Cyber Threat Intelligence (CTI) Sharing:** PRIVATEER incorporates mechanisms to enable the sharing of CTI in a way that preserves privacy through searchable encryption.

PRIVATEER's innovations will be validated in a real-world test network, where the framework will be deployed and tested against relevant use case scenarios. These security mechanisms complement existing 5G/6G security standards and aim to establish a privacy-first security solution that integrates seamlessly into the evolving 6G landscape.

2.1.5.5 RIGOUROUS

The RIGOUROUS [30] is a phase 1 SNS project, which focuses on addressing the critical cybersecurity, trust, and privacy risks that threaten networks, devices, computing infrastructures, and next-generation services in the 6G era. Its mission is to develop a holistic, AI-driven framework capable of dynamically responding to evolving threats across all orchestration layers and network functions.

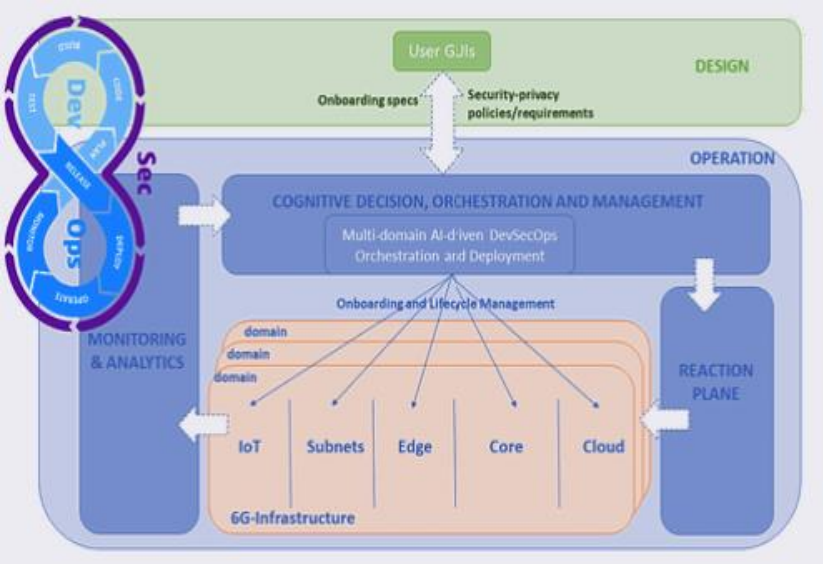

Figure 9: RIGOUROUS architecture [30]

The project introduces a smart service framework designed (Figure 9) to ensure secure, trusted, and privacy-preserving environments, supporting the next generation of 6G services along the full device-edge-cloud continuum across heterogeneous, multi-domain networks. This approach spans the entire DevOps lifecycle, covering:

- Service Onboarding to Day-2 Operations: Including anomaly detection, policy enforcement, and mitigation strategies at various levels (physical and cyber).
- AI-Governed Mechanisms: Aiming to meet security-related requirements and incorporating human-in-the-loop concepts throughout the orchestration process.

RIGOUROUS aims to deliver several key innovations:

1. A Holistic Smart Service Framework for lifecycle management in the IoT-Edge-Cloud continuum.
2. Human-Centric Design with integration of the human factor across the orchestration process.
3. Model-Based and AI-Driven Automated Security Orchestration and Trust Management.
4. Advanced AI-Driven Anomaly Detection and Mitigation Strategies.
5. Demonstration of use cases in real-world operational environments to validate the framework.

The project ultimately seeks to bring intelligence and automation to smart and secure orchestration in future 6G systems.

2.1.5.6 ELASTIC

The ELASTIC [31] is a phase 2 SNS project, which focuses on advancing 6G service orchestration through the integration of cutting-edge cloud-native technologies. It aims to provide a flexible, scalable, and secure framework for managing complex services and workloads across the entire 6G network architecture, from core infrastructure to edge devices.

Key objectives of ELASTIC include:

1. Innovative Service Orchestration: ELASTIC leverages WebAssembly (Wasm), Function-as-a-Service (FaaS), and edge IoT orchestration to enable serverless deployment, dynamic scaling, and optimization of services across 6G networks. This ensures seamless, efficient service delivery and processing of large datasets.

2. Privacy-Preserving AI and Federated Learning: The framework supports multi-party AI and federated machine learning, ensuring data privacy and authenticity across diverse service contexts.

3. Confidential Computing and Security Enhancements: ELASTIC addresses 6G security challenges using Trusted Execution Environments (TEEs) and integrates Confidential Computing with strong security policies. The project also includes research on eBPF/XDP for enhanced security monitoring and WASM security properties.

4. Comprehensive Infrastructure Development: ELASTIC will create an open-source framework and cross-platform drivers, offering real-time observability and low-latency monitoring of workloads. The project emphasizes contributions to open-source communities such as Linux and Kubernetes.

2.1.5.7 ITRUST6G

The iTrust6G [32] project is part of the phase 2 SNS projects and envisions a cutting-edge security architecture tailored for the advanced use cases and applications of 6G networks. By integrating innovative concepts, iTrust6G proposes a unified and intelligent security framework for both distributed network and cloud domains, ensuring flexible and cost-efficient deployment in 6G infrastructures. The project emphasizes a Zero-Trust (ZT) based architecture, designed to address critical 6G security challenges through the following key elements:

1. ZT Security Framework: iTrust6G introduces a software-defined zero-trust architecture for 6G, focusing on enhancing trustworthiness at various levels by ensuring secure, compliant, and observable systems.

2. AI/ML-Driven Security Enhancements: AI and machine learning algorithms are employed for threat detection, handling, and mitigation, improving the accuracy and efficiency of security management. These technologies also enable automated security and dynamic orchestration of services, contributing to more robust and scalable security procedures.

3. Performance Optimization: The project seeks to improve key metrics such as:
    3.1. Trust determination (e.g., faster sharing of cyber-threat intelligence and quicker vulnerability assessments).
    3.2. Security procedure performance (e.g., reducing the time to detect and respond to threats).

3.3. Resource efficiency (e.g., minimizing the performance overhead introduced by programmable security models).

4. Comprehensive Trust Management: iTrust6G leverages trust assessments and AI/ML algorithms to ensure asset compliance and explainable security policies. This approach increases observability across assets and aligns security procedures with 6G requirements.

5. Unified Beyond 5G/6G Vision: iTrust6G's architecture integrates enablers that connect the human, physical, and digital worlds, promoting an intelligent security fabric to support the next generation of highly distributed, secure 6G services and applications.

2.1.5.8 NATWORK

The NATWORK [33] is a phase 2 SNS project, which aims to lay the foundations for a bio-inspired, AI-driven cybersecurity and resilience framework for 6G networks, focusing on economic feasibility, energy efficiency, and self-adaptability. Taking a holistic approach, NATWORK addresses the security challenges of intelligent networking and services across multiple sectors by developing a novel self-adaptive security mechanism based on bio-mimicry principles, enhancing the malleability and resilience of future 6G ecosystems while ensuring service continuity at low energy costs.

The key components and goals of NATWORK are:

1. AI-Leveraged Self-Adaptive Security Mechanism:
   - NATWORK introduces bio-mimicry principles to develop AI-based security mechanisms that enable 6G networks to autonomously adjust and self-regulate to provide secure services while complying with SLAs.
   - The focus is on self-resilience, enabling systems to respond to security threats dynamically and continuously.
2. Secure Federated Learning Architecture:
   - NATWORK's architecture relies on decentralized AI models embedded in the 6G network's physical layer, smart Edge Network Interface Cards (NICs), and RAN devices. These devices are equipped with P4-programmable data planes and DPU acceleration, allowing for local feature extraction at wire-speed and AI model training.
   - This architecture promotes privacy-preserving learning and distributed threat detection.
3. Energy Efficiency and Sustainability:
   - **Net Zero AI** and **energy-efficient security** solutions are at the core of the project, ensuring that security mechanisms operate sustainably while maintaining performance and resilience.
   - The goal is to minimize the energy costs of security operations in 6G environments.
4. Key Security Challenges Addressed:
   - **Moving Target Defense**: NATWORK introduces adaptive responses to dynamically changing threats.
   - **Deep Control Flow Monitoring**: Enhancing detection capabilities for new types of attacks.

- **In-Network Security**: Ensuring a continuum of security for novel in-network operations, supporting secure distributed computing and payload deployment.

5. Decentralized and Modular Orchestration:
   - NATWORK supports decentralized orchestration services, allowing for both horizontal integration (of similar functions using different technologies) and vertical integration (across different domains).
   - APIs will be developed for seamless integration between microservices for security, management, and self-governance.
6. Rapid Verification and Testing: NATWORK emphasizes rapid verification through automated unit testing of its microservice components in simulation environments and lab-based testbeds. This enables fast bug fixing, design adjustments, and performance analysis.

In summary, NATWORK aims to establish a bio-inspired, AI-driven cybersecurity framework for 6G networks that is economically viable, energy-efficient, and capable of dynamically adapting to evolving threats, all while maintaining service continuity and security at every level of the network ecosystem.

2.1.5.9 ROBUST-6G

ROBUST-6G [34] is a phase 2 SNS project which envisions pioneering data-driven, AI/ML-based security solutions that address the evolving cybersecurity challenges within the dynamic and complex 6G cyber-physical continuum. With the anticipated 6G services set to revolutionize communication, ROBUST-6G focuses on creating a holistic, autonomous, and sustainable security framework that ensures robust protection, privacy, and transparency in AI-driven 6G networks.

The key objectives of the project are:

1. **AI/ML-Based Security Solutions:** Develop advanced AI/ML-driven security functionalities that respond to the dynamic threat landscape in 6G networks and services. This includes detecting and mitigating physical layer threats and ensuring resilient and secure network operation.
2. **Zero-Touch Security Management:** Implement fully autonomous, zero-touch security management and resource allocation functionalities to tackle the increasing complexity of 6G networks. This will enable automatic threat detection and response without human intervention.
3. **Protection of AI/ML Systems:** Safeguard AI/ML systems from security attacks, ensuring the privacy of individuals whose data fuels these systems. ROBUST-6G also aims to address concerns about the integrity and reliability of AI models.
4. **Privacy-Preserving and Explainable AI:** Focus on privacy-preserving distributed intelligence while enhancing the transparency and explainability of AI/ML systems. This ensures that security measures are not only effective but also understandable and trustworthy.
5. **Energy-Efficient and Sustainable AI:** Emphasize the development of green AI methodologies that optimize energy efficiency in 6G networks, balancing computational demands with sustainability. ROBUST-6G aims to ensure that security functionalities contribute to a low-energy, environmentally friendly future.

6. **Leveraging Diverse Data Sources:** Utilize multiple data sources (e.g., sensing, positioning, authentication) combined with AI/ML methodologies to detect and mitigate various security threats, including those at the physical layer of 6G networks.

ROBUST-6G's vision is to be at the forefront of securing the 6G ecosystem by building a resilient and intelligent security architecture that integrates AI functionalities across heterogeneous environments. This architecture will address complex security needs, from automated threat detection to privacy protection, while promoting sustainable and energy-efficient practices.

2.2 ENHANCING THE STATE OF THE ART WITH SAFE-6G

SAFE-6G aims to go beyond the existing standards and projects by creating a native trustworthiness framework that integrates security, privacy, resilience, and reliability as the core features of a user-centric 6G system. SAFE-6G's Cognitive Coordinator, Trust Functions, and AI-enabled orchestration differentiate it significantly, addressing current gaps in flexibility, automation, and distributed trustworthiness management.

2.2.1 STANDARDIZATION AND SPECIFICATION

While ISO/IEC 27001 and 3GPP TS 33.501 provide foundational security controls and architectures for centralized 5G networks, SAFE-6G leverages these standards by advancing protocol centric security towards a trust-based approach. Unlike the protocol-centric trust in 5G, SAFE-6G incorporates user-specific, real-time trust configurations through AI, ensuring security, reliability, and resilience as user-defined, adaptive features. SAFE-6G's use of blockchain for verifiable credentials also enhances 3GPP's security protocols by integrating decentralized trust for identity management.

On the other hand, ETSI TS 103 305 focuses on NFV security for 5G networks. This focus aligns with SAFE-6G's architecture. SAFE-6G extends this by incorporating distributed edge-cloud continuum support and intent-based networking, which are critical for adapting NFV to a decentralized 6G framework.

2.2.2 ADVANCES OVER SNS PROJECTS

SAFE-6G advances on several SNS projects by implementing a highly user-centric, intent-driven approach to network trustworthiness, integrating dynamic privacy, security, and reliability measures tailored to each user's unique needs at the current time. While projects like CONFIDENTIAL6G focus on privacy as a static function, SAFE-6G adapts privacy handling dynamically, personalized through its Cognitive Coordinator to align with individual data governance requirements. Similarly, SAFE-6G enhances the zero-touch orchestration models explored by projects like DESIRE6G and HORSE by using intent-based controls, allowing the Cognitive Coordinator to translate user intents into precise trust-based actions, creating adaptable, resilient network behaviours across a distributed cloud continuum. SAFE-6G's AI-driven resilience mechanisms build on the AI anomaly detection foundations of ROBUST-6G and RIGOUROUS, adding real-time, user-specific threat management through a distributed network of AI agents that support continuous, proactive security across diverse network conditions.

2.2.3 CLOUD NATIVE AND ORCHESTRATION

SAFE-6Gs AI-enabled orchestration builds on ONAP's orchestration foundation by enabling dynamic user-specific trustworthiness configurations across distributed resources. This supports 6G's anticipated resource demands more efficiently than ONAP's traditional NFV orchestration. More particularly by using Kubernetes for adaptable service deployment across the cloud continuum.

The Cognitive Coordinator and MLOps in SAFE-6G automate model management, ensuring that AI models are continuously trained to meet evolving trustworthiness requirements. This extends XGVela's observability and resiliency features, making them user-centric rather than function-based.

2.2.4 TRUSTWORTHINESS INNOVATIONS IN SAFE-6G

SAFE-6G's separation of User Service Nodes (USNs) and Network Service Nodes (NSNs) enhances user data ownership and control, contrasting with the NF-centric design in 5G. SAFE-6G supports fine-grained control over data, session management, and mobility management per user. Also, SAFE-6G's distributed trustworthiness framework exploits the edge-cloud continuum for real-time user-specific trust configurations, addressing the unique challenges of a decentralized, multi-stakeholder 6G environment. Finally, Cognitive Coordinator, coupled with eXplainable AI (XAI), interprets user intents for trustworthiness, mapping these to the five trust functions: safety, security, privacy, resilience, and reliability. This intent-based cognitive model advances current trustworthiness models by providing explainable, autonomous trustworthiness management.

3 SAFE-6G RATIONALE

The 6G vision for an open, distributed and user-centric evolution of the current SBA core network creates many security issues. The disaggregated heterogeneous cloud continuum (i.e. distributed cloud system with many stakeholders located in different regions, while private, public, or hybrid clouds are considered for the formation of the continuum), in conjunction with softwarization and IT-based infrastructure operations, sets the stage for risks and challenges to trustworthiness in the 6G era.

The existing 5G security architecture is adaptable to a centralized network architecture, and in 5G, trust connections between network parts are created at the protocol level, rather than involving device and network behaviour. In the envisioned 6G ecosystem, trusted connections are critical for all parties involved, extending security and privacy to a more inclusive framework, such as trustworthiness, which should be assured as a native feature. Therefore, the most significant paradigm adjustments in the envisioned user-centric 6G system are the shift from a security-only focus to a broader scope of native trustworthiness, clarifying that the term "trustworthiness" refers to a holistic approach.

On top of the open and distributed 6G core over the edge-cloud continuum, SAFE-6G proposes holistic research approach aimed to design, develop, and validate a 6G-ready native trustworthiness framework by enabling user-centric safety, security, privacy, resilience, reliability functions. By utilizing (X)AI/ML techniques to cognitively coordinate and balance these functions, to optimize the LoTw, which realises the trust requirements and data governance policy that each user/tenant/human-role specifies, a feature that is considered an essential KVI for user centric 6G.

Verification and validation of the proposed SAFE-6G framework will be performed at the Stream C SNS 6G-SANDBOX Athens platform using two Metaverse-based pilots. The immersive applications will be tested considering different 6G system setups, different service flavours and deployments, under various threats and attacks.

The current operator-centric approach of 5G leads in network architectures that must support a wide range of services (e.g., regional or national level). This has resulted in monolithic and generic network entities and Network Functions (NFs), such as access management functions (AMFs) and user plane functions (UPFs) in 5G, which are expected to serve a large number of end users at the same time without being able to adapt their functionalities to the user specific needs and requirements. So, it is acceptable to assume that a centralized and one-for-all architecture is a natural outcome of the NF-centric design principle.

The SAFE-6G moves beyond the current NF-centric core network towards a user-centric evolution of the B5G/6G system over the recently researched edge-cloud-continuum, which is expected to be the primary option as infrastructure for deploying the softwarised components of a distributed 6G network. Therefore, for 6G to become the human-centric system of systems requires significant architectural redesign based on the user-centric (i.e., per-user perspective), given that the network intrinsically handles the state of each UE or user. A user-centric design is specifically capable of providing to each user a complete instance of a personalised 6G system through a user-specific core-

network synthesis, supporting for example personal data management, policy control, session control, and mobility management per-user.

This revisited architecture should allow users to have and self-manage their own networks, while avoiding the "one-size-fits-all" philosophy to support personalized services. Each user will have a separate network that consolidates all necessary tasks for service delivery thanks to the user-centric architecture.

One might argue that the currently supported "per-user" slicing mechanisms of the existing NF-centric 5G core could accomplish this user-centric vision of 6G. However, this is not feasible, since the existing slicing techniques of the 5G system is built on the "NF-focused" fundamental underlying architecture, where each user may feel like having his/her own dedicated network services, but some centrally located monolithic NFs serve multiple users and UEs creating this user-centricity "illusion", while in practice scaling and performance issues remain.

This paradigm shift from "Network Function-focus" to "user-focus" is essential to manage the growing complexity and demand of modern monolithic network services. These services, both physical and virtual, could become bottlenecks due to the increasing number of connected users and devices. To overcome this, a user-centric architecture will allow users to participate in network service creation and operation, while also granting them full control over data ownership [35] .

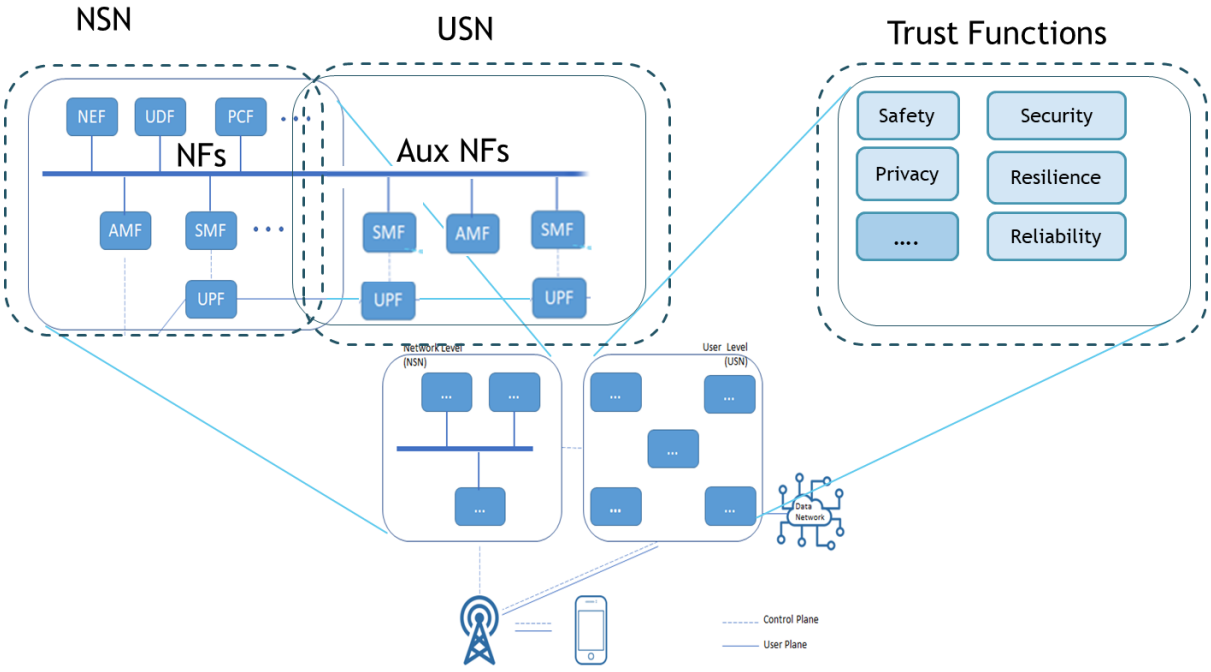

Figure 10: Core network evolution with NSN, USN and Trust Functions

SAFE-6G design is based on the concept of USNs and NSNs. This design ensures that each user receives a personalized instance of the 6G system, tailored to their service requirements and trustworthiness levels. NSN plane includes practically the typical 5G core network, being non-user specific/centric and without special treatment per-user. The USNs playing a pivotal role in the SAFE-6G concept. They implement customized services and policies at the individual user/tenant level. This approach makes

sure that a wide range of service needs are satisfied for each user, ensuring personalized data management, policy control, session control, and mobility management. USNs can adapt in real-time to the dynamic needs and trustworthiness levels of the users.

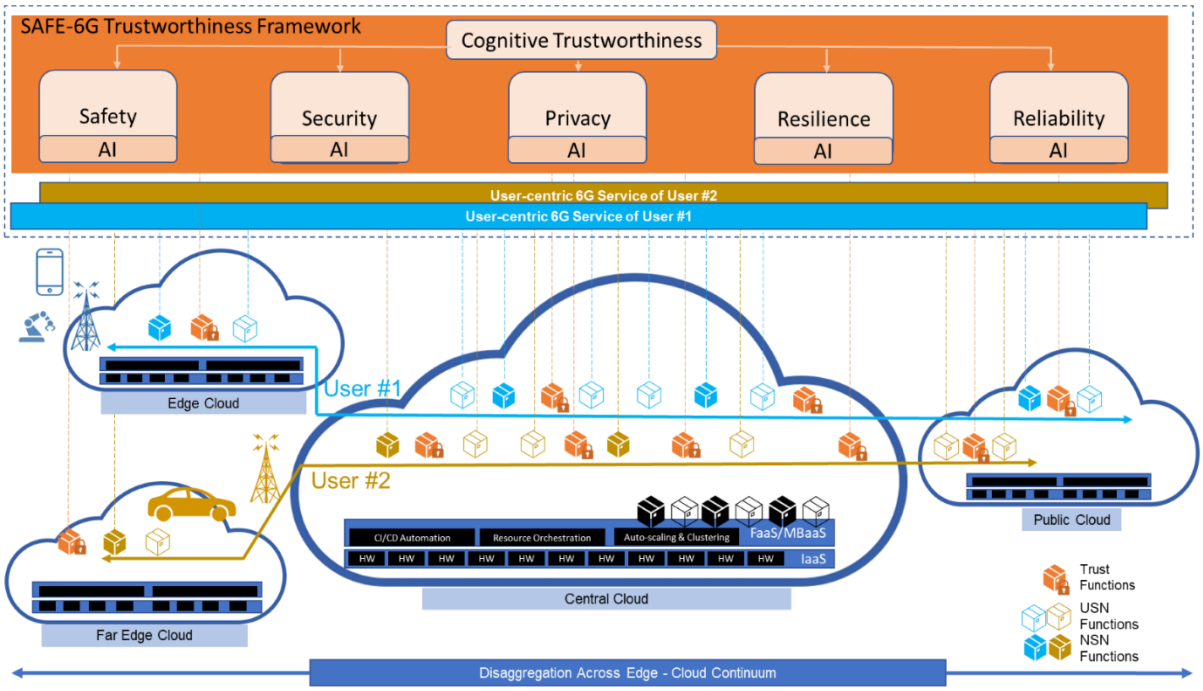

Figure 11: SAFE-6G NSN, USN and Trust Functions placement over the Continuum.

On top of this, in SAFE-6G, native 6G trustworthiness is realised by five user-centric functions: safety, security, privacy, resilience and reliability, which are inter-orchestrated by the Cognitive Coordinator and are deployed as microservices within each user's/tenant's service mesh over the heterogeneous and distributed edge-cloud continuum. The placement decision of each user-centric across the continuum, the balancing among them, their configurations and applied policies are decided by the Cognitive Coordinator that converges data and knowledge from the AI-agents across all the layers vertically and all the heterogenous domains of the continuum horizontally that form the distributed 6G system. Thus, the five Trust Functions (i.e. safety, security, privacy, resilience, and reliability functions) complement the USN and NSN functions in the realisation of the trustworthiness level that is required to support the creation of the level of trust that the user requires for using the 6G system for a specific service/application.

3.1 TRUSTWORTHINESS PROVISION

SAFE-6G defines a comprehensive **trustworthiness** framework for 6G that elevates security, privacy, resilience and reliability from add-ons to native architectural properties. Rather than relying solely on protocol-level trust links as in 5G, SAFE-6G adopts a holistic model that accounts for device behaviour, network reliability and data governance across a distributed cloud continuum. The goal is to ensure safety, privacy and continuous service delivery even as threat landscapes evolve and network functions become more disaggregated, and software driven.

As defined in D2.1, the SAFE-6G framework explicitly treats trustworthiness as a system property that generates user trust, and it positions that property as configurable and adaptive to different tenant and application needs. This reframing addresses the new risks introduced by softwarization, multi-stakeholder cloud continua and pervasive AI components in 6G systems.

SAFE-6G combines a security-by-design philosophy with **intent-based trustworthiness**, using AI and ML to translate high-level trust intents into concrete system configurations. The framework balances competing trust dimensions by design, recognizing that improvements in safety, security, privacy, resilience or reliability each carry costs in usability, agility or performance. By making these tradeoffs explicit and automatable, SAFE-6G enables tailored trust profiles at per-tenant granularity.

A central element is the **Level of Trustworthiness (LoTW)**, a dynamic metric that the system continuously optimizes. Cognitive coordination technologies monitor telemetry, model behaviour and policy constraints to maintain an appropriate LoTW for each user, tenant or application in real time.

SAFE-6G clarifies terminology that is often conflated: **trust** is the attitude or requirement a tenant holds toward the system, while **trustworthiness** is the objective property of the system that produces that trust. The project emphasizes that increasing system trustworthiness raises tenant trust levels, and that precise definitions are essential for consistent design, evaluation and communication across standards bodies and industry stakeholders.

This conceptual clarity underpins SAFE-6G's technical workstreams and evaluation criteria, ensuring that metrics, controls and validation methods align with the intended meaning of trustworthiness rather than informal or inconsistent uses of the term.

At the heart of SAFE-6G is an AI/ML assisted Cognitive Coordinator that interprets abstract trust intents and maps them to operational actions across the five trust dimensions: Safety, Security, Privacy, Resilience and Reliability. The coordinator functions as an intent-handling engine: it ingests tenant policies and runtime signals, computes an ideal trustworthy goal state, and orchestrates the transitions needed to reach and maintain that state.

This cognitive layer performs continuous monitoring, prediction and automated remediation, enabling the system to anticipate vulnerabilities, adapt to changing threats and reconcile conflicting trust requirements. The approach treats cognition as an operational capability—learning from experience and telemetry to improve mappings between intent and configuration over time.

SAFE-6G advances a user-centric model in which trustworthiness is not static but configurable per tenant and per service. The framework supports fine-grained service differentiation by allowing the core network functions and orchestration policies to be tuned to each tenant's required LoTW. This enables dedicated, per-user service profiles that reflect different risk appetites and regulatory constraints.

By making trustworthiness adaptive, SAFE-6G aims to deliver tailored guarantees without imposing a one-size-fits-all posture. The design explicitly accounts for the operational costs of trust measures and

provides mechanisms to trade off those costs against required guarantees in a principled, automated way.

SAFE-6G will validate its concepts through Metaverse-based pilots on the Stream C SNS 6G-SANDBOX platform in Athens, exercising the framework across diverse configurations, service deployments and threat scenarios. These pilots will test the Cognitive Coordinator, LoTW optimization, and the system's ability to maintain trustworthiness under realistic operational stress and adversarial conditions.

If successful, SAFE-6G will provide a practical blueprint for embedding trustworthiness into 6G architectures, deliver evaluation artifacts and deployment patterns, and inform standards and industry practice by demonstrating how AI-assisted, intent-driven coordination can operationalize safety, privacy, resilience and reliability at scale.

4 SAFE-6G Reference Architecture

In this section, the SAFE-6G reference architecture is introduced, as shown in Figure 13. The following subsections will provide a detailed analysis of the various views/perspectives (e.g., user perspective, data perspective) and components of the SAFE-6G User-Centric Cognitive Framework. This exploration will establish a comprehensive understanding of the architecture's foundational elements and its role in supporting the SAFE-6G vision.

This section introduces different views of the SAFE-6G architecture, aiming at creating a complete and multi-dimensional perspective. Additionally, in the deliverables of WP6, additional deployments blueprints will be introduced, referring to the two use-cases that will be deployed for validating the SAFE-6G architecture, elaborating more on the XR/Metaverse service deployment over the SAFE-6G infrastructure, supporting further the replicability across similar vertical industries.

For facilitating the readability of the deliverable, we are providing hereby a traceability matrix between the architectural components of SAFE-6G and the respective section of D2.1 that reports the respective requirements.

SAFE-6G Components	***D2.1 Section/Requirements category***
Chatbot Requirements	Section 5.2
Cognitive Coordinator Requirements	Section 5.3
Safety Function Requirements	Section 5.4
Security Function Requirements	Section 5.5
Privacy Function Requirements	Section 5.6
Resilience Function Requirements	Section 5.7
Reliability Function Requirements	Section 5.8
MLOPS Framework Requirements	Section 5.9
6G Core requirements Requirements	Section 5.10
Continuum Requirements Requirements	Section 5.11

Table 2: Traceability between SAFE-6G components and the corresponding D2.1 requirement sections.

High level view:

Provision of trustworthiness from a 6G system to a user is a characteristic that will become essential in 6G networks. Despite the discrete meaning and scope of trustworthiness, it is still observed that it is usually misused as trust. As previously mentioned, trust is an attitude that a tenant has towards a 6G system. In contrast, trustworthiness is a system property that creates trust to the 6G tenant/user. A user/tenant trusts (or requires a specific level of trust from) a 6G system, because the 6G system is trustworthy. In other words, the trustworthiness of a 6G system contributes to building the trust level of the tenant/user of the specific system. Thus, the more trustworthy the 6G system is, the higher the trust level of the tenant/user will be [36] .

A realistic solution to this trustworthiness challenge must recognize that all security measures (i.e., safety, security, privacy, resilience and reliability) come at a cost in terms of usability, agility, or

swiftness. As a result, the envisioned trustworthiness framework should provide a balance between the various security measures by dealing with a security-by-design approach, as well as a wide range of themes, such as the trust model and the application of new cognitive coordination technologies (e.g., intent-based trustworthiness, based on AI/ML techniques).

This impact of trustworthiness measures on the usability of the system by the tenant may impact the perceived level of trust if the user is not fully understanding the reason that the system takes the actions that are affecting them. Explainability is often viewed as an effective way to build trust among stakeholders. If users have a better understanding of the process by which the system generates its outputs and the explanation provided for a particular result aligns with their preconceptions of what constitutes a proper decision, then the level of trust of the system has been improved. The literature does, in fact, frequently link explainability to trust [37] ,[38] , and many researchers—at least tacitly—assume that explainability and trust are strongly related[39] ,[40] . This relationship is known as the Explainability-Trust-Hypothesis, which states that "Explainability is a suitable means for facilitating trust in a stakeholder"[36] .

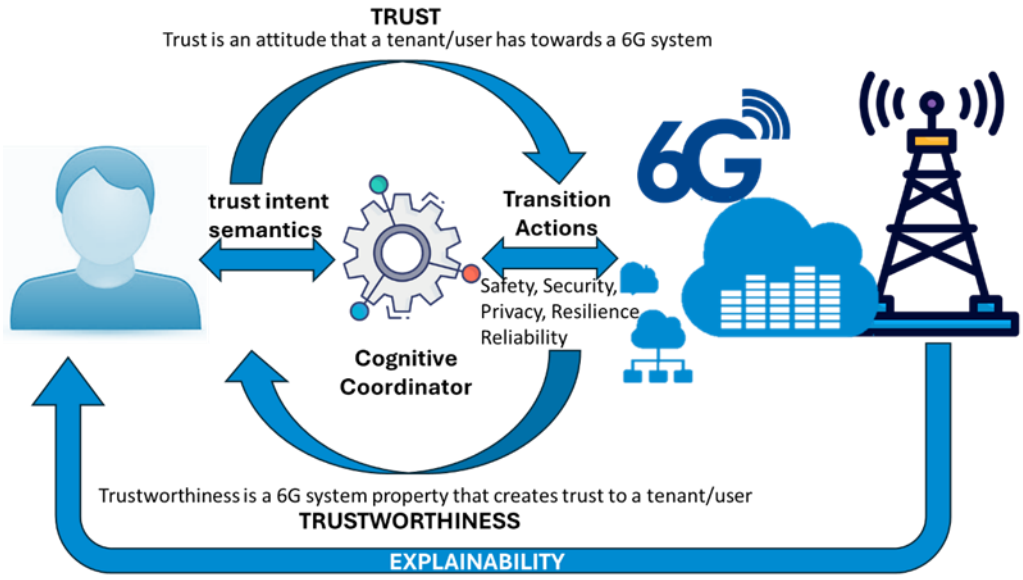

Figure 12: User-centric 6G trustworthiness with explainability feedback.

Among the various explainability tools, SAFE-6G as a modern system has selected in its design principles the use of XAI [41] to complement the operation of the Cognitive Coordinator and contribute to the improvement of the tenant's/user's Level of Trust over the Trustworthy 6G System. As evidenced by the "right to explanation" outlined in the General Data Protection Regulation (GDPR) and the European Commission's (EC) Technical Study on "Ethics recommendations for trustworthy AI"[42] , trustworthiness has become crucial for both users and governmental organizations. They claim that explainability is a crucial element of trustworthiness. As a result, XAI, or an AI "that creates information or reasoning to make its working obvious or easy to understand," is receiving more and more interest from both industry and academia. In this context, two strategies for achieving explainability can be identified: The adoption of post-hoc explainability techniques (i.e., the

"explaining black-box" strategy) and the design of inherently interpretable models (i.e., "transparent box design" strategy). These approaches allow to understand the model behaviour and can be integrated in the model training or being applied as post-hoc approaches, after ML training. SAFE-6G will focus on the first method, because it allows to add the interpretability layer without changing the model training – which is performed within the MLOps lifecycle, using Global XAI methods and Local XAI methods, which will be further analysed in WP4.

Although SAFE-6G does not explicitly target large-scale deployments, since the project focuses on two proof-of-concept scenarios, the architecture still considers future extensibility. While scalability is not a primary design driver, WP3, WP4, and WP5 include plans to explore multi-tenancy mechanisms that would allow the system to support broader deployments if needed. This ensures that the design remains open to future scaling requirements without exceeding the scope of the current pilots.

Summarizing the previously mentioned design principles for SAFE-6G and considering the building blocks of a distributed, open and user-centric 6G system, namely i) the Continuum Plane, ii) the Core Openness Plane and iii) the User-centric-App Plane, a high-level view of SAFE-6G ecosystem is hereby provided, reflecting the high-level operation of SAFE-6G and illustrating the main building domains.

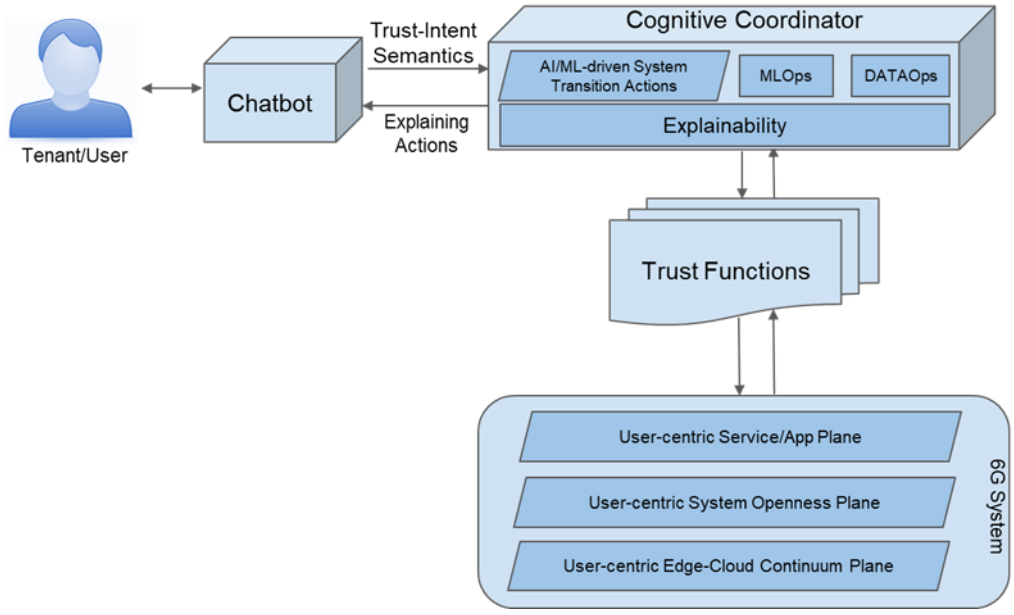

Figure 13: High level view of SAFE-6G Reference Architecture

A brief description of the main building blocks/domains that are derived from the design principles of SAFE-6G and are illustrated in the SAFE-6G blueprint are summarised below:

Chatbot: Allows users to interact with the system to request the LoTw needed through the SAFE-6G Cognitive Coordinator, which extracts the LoTw across the five SAFE-6G user-centric Trust Functions.

Cognitive Coordinator: The SAFE-6G cognitive coordinator is an intent-handling component that comprehends trust intent semantics from the user through the AI Chatbot, calculates the requested LoT, and coordinates the five SAFE-6G user-centric functions to transit the 6G system into this trustworthy state.

Trust Functions (TFs): The TFs are intended to meet particular Level of Trust of a user by increasing the system's trustworthiness and complementing the USN and NSN functions of the 6G core network. More specifically, the TFs include the design and the development of the five AI-assisted user-centric functions. It consists of safety, security, privacy, resilience and reliability. The purpose of those functions is the LoTw provision for a user-centric distributed 6G ecosystem. The functions will consider all the lifecycle phases of the USN and NSN functions including the phases of before service deployment, during service deployment (operation), and after service deployment, to create an adaptive and scalable trustworthiness mechanism. SAFE-6G organises trustworthiness around five interdependent functions—Safety, Security, Privacy, Resilience, and Reliability—that together ensure secure, user-centric 6G operation. Safety enforces strict isolation and placement policies across the cloud continuum so that USN and NSN workloads are protected from unauthorized access and cross-tenant interference. Security adopts a zero-trust posture reinforced by blockchain-backed verifiable credentials and tokenized actions, providing strong authentication, immutable audit trails and lifecycle protection for sensitive data. Privacy is embedded by design: users express preferences via an AI chatbot, and the Cognitive Coordinator translates those intents into placement, access and processing constraints that the AI orchestrators enforce. Resilience leverages AI-driven, intent-based operations to detect, adapt and remediate faults or attacks automatically, enabling graceful degradation and rapid recovery. Reliability is achieved through continuous service and reliability profiling: telemetry trains ML models to recognise anomalies and malicious behaviour, with deployed models providing real-time detection and mitigation. Together these functions form a cohesive trust fabric that maps user intent into measurable, enforceable controls across the 6G ecosystem.

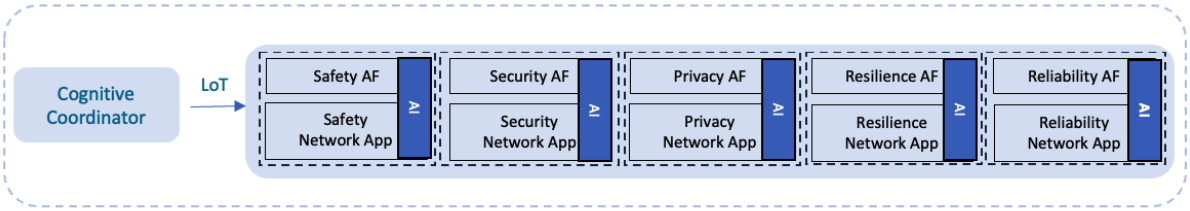

Figure 14: SAFE-6G user-centric trustworthiness functions triggered by Cognitive Coordinator

6G System Planes: The user-centric, open and distributed 6G system in SAFE-6G is represented by three planes, namely i) the User-centric Service/App Plane, ii) the User-centric System Openness Plane, and iii) the user-centric edge-cloud continuum plane.

The user-centric Service/App plane refers to the user-centric applications that are deployed and personalised on top of the 6G-system and the continuum, supporting also the capability to interact with the 6G system through exposed APIs, as it is foreseen in 3GPP SA6 for the vertical applications. The user-centric system openness plane refers to the exposure and programmable capabilities of the core network via standardised APIs that are published via CAPIF framework. Finally, the user- centric edge-cloud-continuum plane refers to the continuum platforms that can support the distributed deployment of applications, but also network functions as well, in order to optimize the user experience by bringing specific functionalities closer to the user. In SAFE-6G it is important to emphasize that openness of the 6G system is considered as an enabling technology for realizing the user-centric provision, which means that each one of the three planes will be exposing their

capabilities through APIs via CAPIF-framework to the SAFE-6G system (i.e. Trust Functions and Cognitive Coordinator) in order to be possible their adaptation to the appropriate configuration state that reflect the trustworthiness level needed for a specific user.

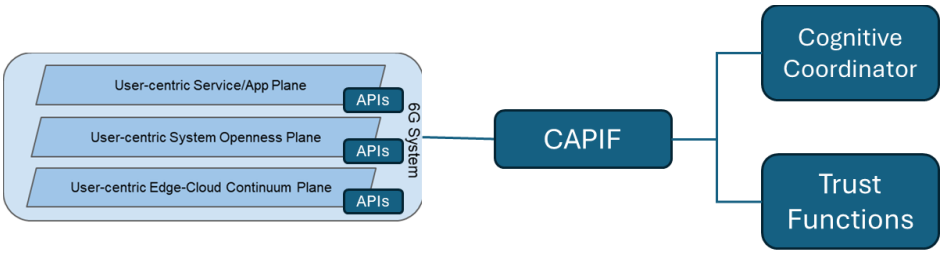

Figure 15: Integration of SAFE-6G components with 6G system's planes via CAPIF.

MLOps: The MLOps component of the SAFE-6G framework will deploy and efficiently maintain AI models in production. It will be responsible for the continuous training and evaluation process of the AI models distributed across the system and that are part of the SAFE-6G framework, and will consist of a set of interfaces, models and libraries.

DataOps: DataOps component will contribute on feeding MLOPs with data. This module offers data processed automatically so it can be leveraged by the different components of the framework. This data can be of two types: real and simulated.

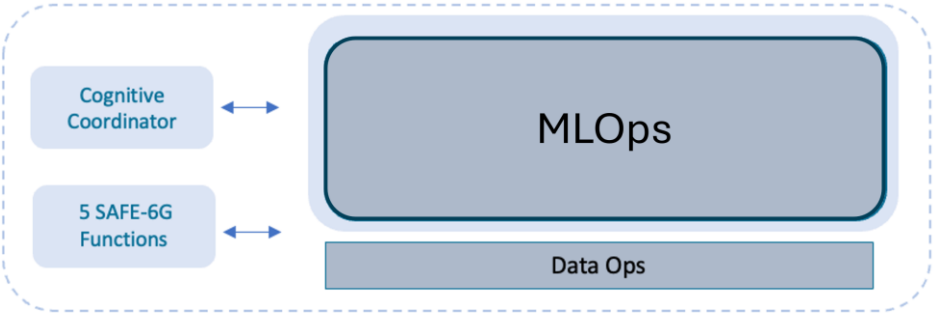

Figure 16: MLOps and Data Ops as the main AI lifecycle enablers in SAFE-6G.

4.1 FUNCTIONAL VIEW

As the 6G network's applicability grows into new services and business contexts, the need for far more customized connectivity solutions that are tailored to the individual requirements and become user/human-centric is expected.

The SAFE-6G project introduces a shift from the traditional operator-centric and function-centric core network architecture, which currently manages each UE's or end user's state by maintaining consistent states across various network functions. This new framework is designed to deliver personalized, secure, and reliable services through a suite of advanced functions, each contributing to a holistic, user-focused network experience.

SAFE-6G reimagines this framework with a user-centric approach, leveraging the edge cloud continuum to create a more personalized, human-centric 6G system. To realize the user-centric 6G network, the core network must be architecturally redesigned to reflect the paradigm shift from "Network Function-focus" to "User-focus," allowing users to engage in network service creation and operation while simultaneously providing users complete control over data ownership.

In this new architecture, each user benefits from a dedicated instance of the 6G system that handles their unique data, policies, sessions, and mobility requirements. Moving away from the "one-size-fits-all" model, SAFE-6G empowers users to self-manage their network environments, including the ability to create customized VPNs. Crucially, the functional view of this architecture is underpinned by five Application Functions (AFs) that act as trust functions, ensuring security, integrity, and tailored service provision for each user. This design not only reduces latency but also allows for the creation of network slices customized to individual user needs, while evolving and enhancing current 3GPP interfaces and functions.

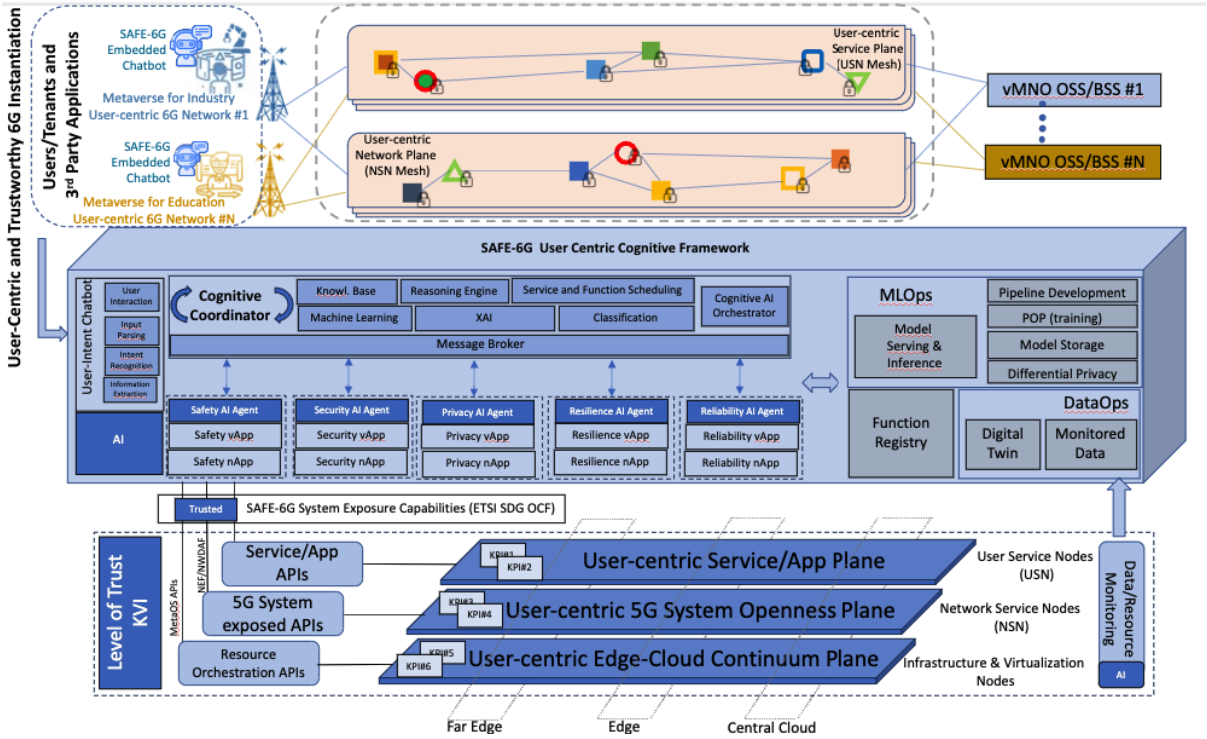

Figure 17: Functional View of SAFE-6G Reference Architecture

These user-centric AFs are inter-orchestrated by the cognitive coordinator and deployed as microservices within each user's/tenant's service mesh over the heterogeneous and distributed edge-cloud continuum. The cognitive coordinator determines the placement of each user-centric attribute within the continuum, balances them, configures them, and applies policies. It does this by integrating data and knowledge from AI agents (shown as blue AI boxes in Figure 17) across all vertical layers and horizontal heterogeneous domains of the continuum, forming the distributed 6G system.

User intent Chatbot. As in 6G trustworthiness preservation is even more challenging given that user traffic may traverse nodes belonging to multiple stakeholders/providers and operators, SAFE-6G

project will treat user's trust as an integral element of user-centric service provision. To this end, a novel data schema will allow the users, as depicted in Figure 17, to express their trustworthiness requirements concerning a new or already deployed service via the auxiliary chatbot interface. These requirements/intents should constitute an input to the SAFE-6G Cognitive Coordinator and will impact primarily the placement and resource management decisions as well as also specific functions to enhance the trustworthiness of the overall ecosystem via their respective AI agents.

The Cognitive Coordinator is a novel element for mapping user/tenant Intents to a LoTw based on the information provided to the chatbot (as shown at middle side of Figure 17). The defined LoTw is then materialized to specific deployments of the user-centric functions for providing to the specific user/tenant with the necessary and appropriate end-to-end safety, security, privacy, resilience, reliability policy. The Cognitive Coordinator aggregates the AI/XAI models of the SAFE-6G framework and all the intelligence and AI agents of the underlying five enabling functions, as well as the necessary programmability intelligence of the resource orchestration of the cloud continuum that realizes the distributed user-centric 6G, and specifically the USN and NSN planes.

The MLOps framework supports and eases the development and execution of AI/ML algorithms across multiple domains. The framework will enable the management of a large part of the lifecycle of AI/ML models, from design, training, and evaluation to deployment in production environments at different domains (far/extreme edge, edge, cloud), allowing their smooth integration into the Cognitive framework. Therefore, SAFE-6G will cope with the challenge of having many closed control loops at every segment of the network, facilitating the AI/ML model development operations in the architecture. The MLOps Framework serving infrastructures will allow serving of multiple models simultaneously, managing both, versioning, and model labelling and separate testing and production environments.

The DataOps module will support MLOPs, with dedicated storage types and databases to support the AI/ML training processes. Part of this module is also **the Network's Digital Twin** (NDT), which extends the notion of digital twins in the domain of networking. In SAFE-6G, NDT will comprise three layers: (i) The first layer will be the physical network layer which is the target real world 6G network operating environment. (ii) The second will be the 6G twin layer which is the core of the NDT, and which integrates the data domain, the model domain, and the management domain. (iii) The third will be the application layer, which enables the two-way interaction between them as it provides the means to exploit the capabilities offered by the 6G twin layer. The whole process will complement the MLOps system of SAFE-6G, which is powered by AI/ML functions that are key to achieving the Cognitive Coordination autonomous feedback loop between the 6G physical network and the SAFE-6G framework. The AI/ML functions produced by the NDT and feed to the MLOps for initiative the cycle of the continuous training and enhancement of their predictions will provide a powerful set of tools for analyzing and understanding complex data. In an NDT, there are typically large amounts of data generated by sensors and other sources, and ML techniques will help to extract insights from this data that would be difficult or impossible to obtain using traditional methods. Therefore, NDT in SAFE-6G will support data generation by simulation for the AI/ML training and inference processes.

The orchestrating features of SAFE-6G system include AI-enabled resource orchestration and service mesh networking aspects. Aiming at coping with the increasing heterogeneity and distributed availability of computing resources, a meta-operating system (Meta-OS) for the computing continuum will manage the virtualized resources and network. It should be mentioned that despite performing a similar function as an ETSI NFV-MANO framework, implementations are not yet compliant with this specification. **To achieve zero-touch resource allocation,** trustworthiness requirements/intents coming from the chatbot will constitute an input to the SAFE-6G Cognitive Coordinator and should impact primarily the placement and resource management decisions of the Meta-OS. Besides, the latter will be in charge of the service mesh features of the framework. Apart from allowing an optimal balancing of workloads among replicas of a given microservice and resilient connectivity, it offers (i) the possibility of enabling/disabling networking policies, (ii) observability features (tracing, metrics, etc.) with very little overheads, and (iii) automatic encryption the inter-node/clusters transmissions.

6G is expected to rely on a continuum of a high number of heterogeneous resources connected through multiple network domains**,** whose boundaries will be blurred, enabling the realization of the distributed NSN and USN planes. The integration of those heterogeneous resources adds a lot of complexity since the so-called far- or extreme-edge may be a very volatile environment, as it may be composed of many different devices based on a wide variety of technologies, including both hardware and software. This leads to much higher complexity from the management perspective, for which a **Resource monitoring module** will be key to:

1. Support predictive orchestration.
2. Increase automation (zero-touch).
3. Feeding trust functions and AI models with relevant historic and contextual data.
4. Improve resiliency enabling more secure and private exchanges between the participating domains.

SAFE-6G introduces five enablers/functions for trustworthiness, namely Safety, Security, Privacy, Resilience and Reliability, whose service placement and provision are coordinated by the Cognitive Coordinator in conjunction with the AI-driven resource and service mesh orchestrators, achieving an advanced and fully automated system (zero-touch) across all the network layers that form the SAFE-6G ecosystem.

Safety within SAFE-6G integrates the Software Defined Perimeter (SDP) paradigm within the User-Centric 6G Packet Core infrastructure. This technology establishes secure, individualized perimeters around critical network services such as nodes, controllers, and data centers. By doing so, SAFE-6G limits exposure to potential security threats and ensures that access is granted only to authorized users based on their specific needs. The SDP approach, grounded in the zero-trust security model, employs micro-segmentation of entitlements, creating a finely tuned security environment that adapts to the user's requirements. This ensures that the network remains resilient against unauthorized access and other security threats, safeguarding the infrastructure while providing a tailored experience for each user.

Security within SAFE-6G is further enhanced using blockchain-based verifiable credentials and the tokenization of user actions. These technologies ensure that all access and interactions within the

network are tightly controlled and verified, reducing the risks associated with identity theft, data breaches, and other cyber threats. The decentralized nature of blockchain provides a transparent and tamper-resistant system, while regular security audits ensure that vulnerabilities are identified and addressed proactively. By integrating these advanced security measures, SAFE-6G creates a robust and trustworthy environment, essential for handling the sensitive and personal data internal to the 6G ecosystem.

Privacy is a fundamental pillar of the SAFE-6G project, seamlessly integrated into its user-centric architecture. The project introduces mechanisms that allow users to define their privacy preferences through intuitive interfaces or APIs, depending on their role within the network. These privacy requirements are fed into the SAFE-6G Cognitive Coordinator, influencing decisions related to resource placement and management by AI-driven orchestrators. By embedding privacy considerations into the core operational processes, SAFE-6G ensures that user data is managed with the utmost confidentiality, aligning with the highest standards of user trust and control.

Resilience within the SAFE-6G network is achieved through an intent-based approach that utilizes a sophisticated ontology of intent, built on knowledge management and semantic modelling techniques. The resilient function is designed to interpret user intents and translate them into actionable network operations that optimize service delivery. Machine learning and a knowledge base structured using the RDF Schema enable the network to adapt dynamically to changing conditions and user needs. This adaptability ensures that the network remains robust and capable of delivering continuous, efficient service, even under varying operational demands.

The **Reliability** of the SAFE-6G system is underpinned by comprehensive service and reliability profiling, which involves multi-layer data collection across physical, virtual, and application layers. This data feeds into federated learning-based mechanisms that ensure data privacy while enabling advanced machine learning capabilities. By continuously monitoring service performance and simulating various operational conditions, SAFE-6G can detect and respond to potential threats in real time. This approach ensures that the network maintains high reliability and consistent service quality, even in the face of diverse challenges such as DDoS attacks or other forms of intrusion.

Collectively, these functions form a transformative architecture that not only redefines 6G network service provision but also sets a new benchmark for user-centered innovation in telecommunications. SAFE-6G's seamless integration of cutting-edge technologies like SDP, blockchain, and AI underscores its commitment to creating a network that is not only more secure and reliable but also fundamentally more aligned with the needs and expectations of its users.

4.2 Process view & Data View

The **Process view** describes the flow of the system (Figure 18) through a user perspective and goes as follows:

- The **tenant/user** communicates with the SAFE-6G system through the **chatbot**, which acts as an interface between them. More specifically, the **tenant/user's** intents/needs related to safety, security, privacy, resilience, and reliability are feeding the **chatbot** in text form.

- The **Chatbot** based on an AI-assisted NLP procedure, converts the **tenant/user's** intents/needs into semantics, e.g. it classifies each sentence to a **TF**, and communicates them to the **Cognitive Coordinator**.
- The **Cognitive Coordinator** calculates the nLoT value based on the semantics provided by the chatbot, e.g. total number of sentences, number of sentences per TF.
- By exploiting the underlying knowledge base (user info, system resources), the Cognitive Coordinator refines nLoT into calibrated, deployable trust scores (cLoT) for each of the five Trust Functions (TFs). These cLoT scores (in the range 0–100) are directly consumed by the TFs, which then execute the corresponding actions (e.g., activation or adaptation of specific ML methods and policies) to materialise the requested Level of Trust. The SAFE-6G **Orchestrator** deploys each **TF's** instance through the TF **AI Agent**. It communicates also with the MLOps to get the pre-trained models based on the score provided by each function.
- The **MLOps Framework** provides and ensures that pre-trained ML models used by the SAFE-6G **Orchestrator** are up-to-date and operate correctly, e.g. converge to sufficient accuracy.
- **The CAPIF** provides the framework for accessing the APIs so that the SAFE-6G components can communicate and integrate seamlessly. In each **TF**, on start-up, **vApp** uses the **CAPIF** to communicate with the **6G System's planes** to request specific metrics that will be exposed via the monitoring framework for the ML inference phase. These data are being received by **TF's NetworkApp**.
- The deployment of **TF** is in a loop, and in the end, the **AI Agent** reports/alerts the **Cognitive Coordinator,** e.g. by providing the actual feasible score, other alerts, etc. Furthermore, in this loop, the **vApp** communicates with the **AI Agent** and the **NetworkApp** until it finishes its operations.
- The **Monitoring Framework** continuously checks the performance of all 6G components, ensuring they meet the required standards. This approach provides flexibility to add new functions, and each function can request the data and metrics needed to reach the LoT, as there is freedom for each function to select and exploit the available data it wishes.
- The outcome of the TF deployments is reported back to Cognitive Coordinator who communicates with the **xAI module** to get some level of explainability of the overall performance (cLoT, deployment outcome) and then it offers this explainability along with the cLoT to the **Chatbot**.
- Finally, the **Chatbot** informs the **tenant/user** about the cLoT.

All these components operate within the different planes of the 6G system, ensuring efficient and secure service delivery.

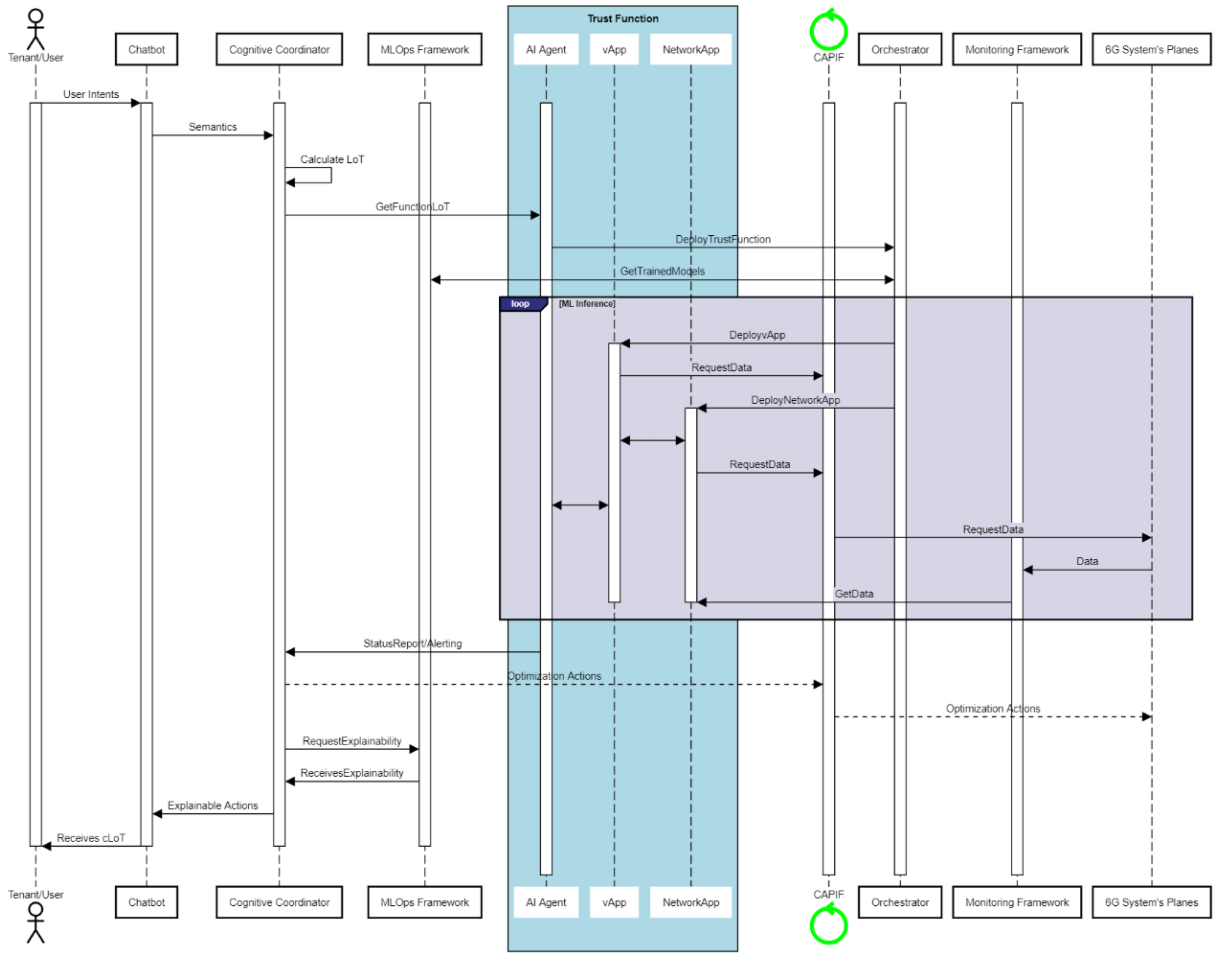

Figure 18: Sequence diagram illustrating the main processes within the SAFE-6G architecture.

The **Data view** in SAFE-6G Project refers to the workflow and lifecycle of the data across the continuum in 6G system.

The workflow of the data treatment will be as follows:

1. The user will interact with the SAFE-6G system to request a LoTw through the chatbot.
2. The chatbot, by making use of the cognitive layer, will make service-driven questions to the user. The responses from the user are taken as input data.
3. These data will be processed within the Cognitive Coordinator Layer. In such layer, the different modules (regression and reasoning components) will convert, quantify, and structure the inputs from the user.
4. Then, the Cognitive Coordinator will calculate the LoTw across the five user-centric functions (trust intent semantics).
5. Once the LoTw is calculated, the trust functions interacting with AI agents will determine necessary actions to apply in 6G system to increase the trustworthiness.

The SAFE-6G systems encompasses three planes to monitor and collect network and user data:

a. User-centric Service/App plane refers to data related to the applications on top of 6G system exposed through APIs (3GPP SA6).
b. User-centric System Openness Plane refers to data related to the core network.
c. User-centric Edge-Cloud Continuum Plane refers to data of applications distributed in the continuum (central cloud, edge, far-edge) as well as the network functions.

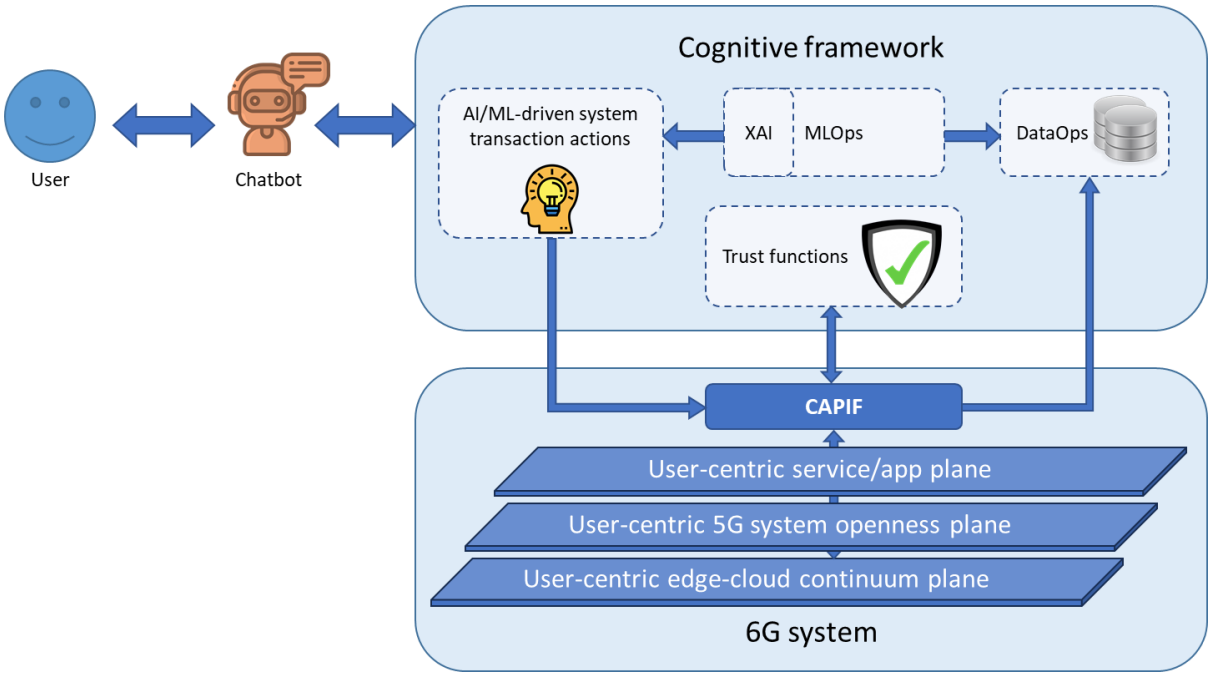

Figure 19: Main flows of data in the ecosystem.

All the relevant data managed by either the cognitive coordination layer, the trust functions or the 6G system will be monitored, collected, and provided to the MLOps via the DataOps module, responsible for cleaning them up to create a useful dataset (labelled, with no inconsistencies and/or duplicated data, etc.). After having gone through the MLOPs module, the data, alongside with a trained ML model, will be provided to the XAI module to give the user explainability about the AI models to be applied for achieving the requested LoTw.

4.3 Deployment view

SAFE-6G is being designed considering two complementary paradigms: The Edge Cloud computing continuum and Cloud Native, which are promoted by the EC and the telco industry, respectively. These paradigms come with a set of requirements and best practices that SAFE-6G must consider, at design, development, and deployment stages. This view is related to the last of them, representing the process of deploying the SAFE-6G system in the telco ecosystem and making it available and operational for mobile network operators and (ICT, metaverse) application providers.

Implementing the SAFE-6G trustworthiness framework demands as pre-requisites the existence of key functionalities related to virtualization, container orchestration and computing continuum management. To ease the realization of SAFE-6G, this view also considers the provisioning of those functionalities, including the technological selection. Therefore, the deployment view includes two parts: (i) the deployment of the virtualization technologies and the meta-orchestrating system that will handle the orchestration of resources and services across the computing continuum (aerOS is considered in the project, although another could be used if includes features as those indicated in Section Edge cloud Continuum); and (ii) the installation of the rest of the building blocks of the SAFE-6G trustworthiness framework. This view presents the system from an engineer's point of view while deploying, placing, configuring, and interconnecting all the necessary software components on the physical layer needed to ensure that SAFE-6G capabilities are accessible and ready to operationally serve stakeholders according to their specified and designed intentions. Notice that the deployment of other (network/core/application-related) services is covered by the functional view, enabled when the SAFE-6G system is operational.

Starting with the first part, the process starts by **defining the domains** that will compose the SAFE-6G framework, including the entrypoint domain (the one that will host the management portal). This means identifying all the distributed computing environments that will be part of the SAFE-6G computing continuum, including the devices and owner/s involved (also from cloud platforms' resources). The latter is important as they will oversee – or delegate – the next steps: the installation and configuration of the **virtualization technology** (mainly Kubernetes and related systems and plugins) in the Infrastructure Elements (IEs), and the deployment of **aerOS runtime and basic services** over them to realize each domain. These fully documented processes are supported by available tools and scripts, with each domain ending up with an example of deployment as seen in Figure 20

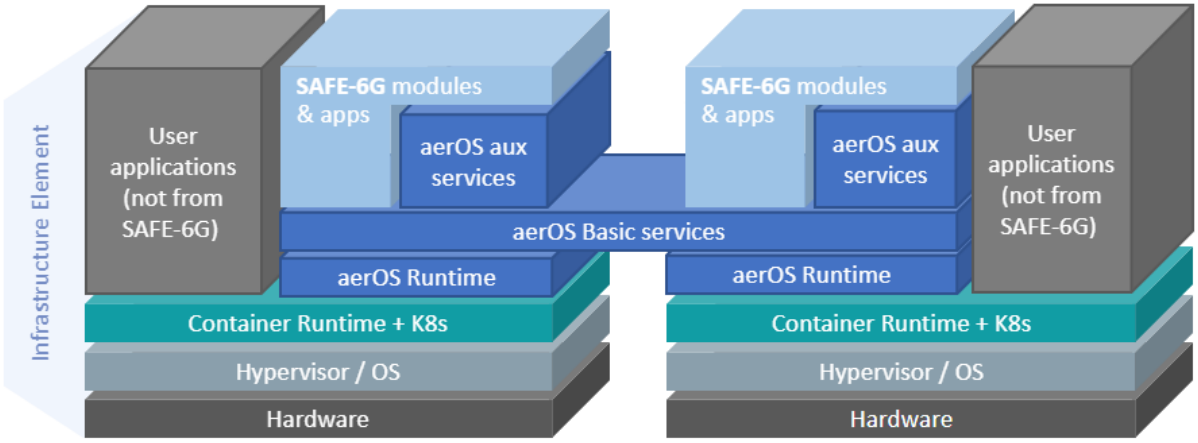

Figure 20: Deployment of aerOS and SAFE-6G components.

Since aerOS is a software product of an external project on top of which some code modifications are needed to adapt it to the telco ecosystem, it is not possible to update it each time the source code is updated. SAFE-6G considers an internal forked version in which the needed changes are implemented periodically, version that can be deployed with the guidelines available (with DevOps and CI/CD tools on the registered domains).

Besides, it should be mentioned that not all aerOS services must be installed in all IEs. The following table depicts the number of instances of its components/services that should be deployed at the continuum, per domain and IE. More information about them can be found in D2.7 of aerOS project [43]

aerOS component/ service	*In the whole continuum*	*In each domain*	*In each IE*
aerOS Management portal	Single Instance (1), in the entrypoint domain	-	-
aerOS Federator	N (as many as domains)	1	-
High-Level Orchestrator (HLO)	N (as many as domains)	1	-
Low-Level Orchestrator (LLO)	$\sum_{i=0}^{N} D^i$ (Total number of LLOs in the continuum (N domains)	D (the multitude of IE container management framework types integrated in the domain)	1 associated LLO
Identity management and authorization (OpenLDAP)	Single Instance (1)	-	-
OpenAPI	N (as many as domains)	1	-
Data fabric	N (as many as domains)	1	-
Self-*	X (one per IE)	X (one per IE)	1

Table 3: Instances of aerOS components per-continuum, domain and IE

Once aerOS is installed in the IEs of a domain, the next step is to **register and enroll the newly created domain to the aerOS continuum**, via the management portal (accessible from the aerOS entrypoint domain). In terms of connectivity, interconnection of aerOS IEs and domains is based on IP/TCP layer but employs a variety of backhauls such as 5G, 4G, internet, WiFi, etc. VPN and tunnelling overlays are used when required.

Continuing with the second part, the remaining SAFE-6G building blocks will be deployed using aerOS orchestration capabilities. These are primarily the services of the SAFE-6G trustworthiness framework and the cellular core/s, of which container images and Helm charts will be available. Particularly, the aerOS' management interfaces (CLI or API) will be leveraged to deploy these services, with guided documentation for the engineers in charge of their deployment. They will have the capability to position them in a specific computing node manually or to let aerOS decide their optimal placement. Regardless of the option implemented, the following modules must be deployed to finalize the deployment of the SAFE-6G framework:

1. Cellular 6G Core.
2. Monitoring probes in the continuum.
3. SAFE-6G Trustworthiness Framework:
    a. MLOPs module.
    b. Cognitive coordination.
    c. Trust functions (i.e., security, privacy, safety, reliability, resiliency).
    d. User-Intent chatbot.

4. OpenCAPIF Framework.

In case new trust functions or additional features are integrated as new modules or within the existing ones, the components involved should be installed or upgraded with the Meta-OS. It is key that they consider the proper version of the exposed interfaces of the systems they communicate with and that these have been previously tested and validated in an integration environment. Once the SAFE-6G system is in place, the chatbot can receive requests with user intents to deploy use case-related services, in this case falling within the scope of the functional view.

4.4 BUSINESS VIEW

The transition from PNFs and VNFs to CNFs is a core aspect of modularity in 6G networks. It directly addresses the industry's drive for increased scalability, flexibility, and operational efficiency. This paradigm shift, grounded in microservices architecture, containerization, and orchestration may have a strong impact in businesses overall. This will be especially relevant in user-centric networks as modularity undeniably comes with tailored experiences by adjusting network resources based on individual user requirements.

With microservices architecture and containerization, companies can efficiently deploy and scale specific services depending on user demands. This tailored approach not only enhances user satisfaction but also enables businesses to offer differentiated, premium service tiers—driving new revenue streams and customer loyalty. On the other hand, containerization and orchestration enable businesses to optimize resource allocation and scale network functions on demand, adjusting in user activity without incurring high costs. This flexible, on-demand scaling reduces operational expenditure (OpEx) while maintaining a high level of user satisfaction, leading to a lower total cost of ownership (TCO) over time.

Trust services are not exempt from this new approach. In this chapter, a threefold business attempt is given to harness the business potential by leveraging the integration of trust functions to elevate service quality and foster customer trust alike.

- Trust as a differentiator: By focusing on resilience, reliability, privacy, safety, and security, businesses can position themselves as trustworthy providers in the competitive 6G landscape. These aspects are especially critical for enterprises and industries that require guaranteed service levels and data protection.
- Tailored Trust Features for Premium Services: Businesses can create tiered service offerings that emphasize these trust factors, such as premium packages that include enhanced resilience or privacy for specific use cases like remote work, healthcare, or IoT applications. Indeed, from the Resilience perspective, additional redundancy can be provisioned by isolating functions into microservices, reducing the impact of a failure in one component on the rest of the system. This enables automated recovery mechanisms that can restart failed containers or shift workloads without manual intervention. This way continuity of service even in the event of disruptions, enhancing customer trust and maintaining service-level agreements (SLAs) can be ensured. From the viewpoint of Reliability, service orchestration plays a key role in dynamically scale CNFs to match demand, ensuring consistent performance

even during peak usage periods. This approach minimizes service degradation and allows businesses to maintain high levels of customer satisfaction. Ensuring enough Privacy, especially at safeguarding users' data via encryption, anonymization, or processing in compliance with local regulations adds extra flexibility for industries and businesses with strict privacy requirements, helping them businesses maintain compliance and build user trust. In terms of Safety, some restrictions to certain services may be applied if the user's retrieved information indicates a potential hazard or high-risk environment. With proper support and automation, businesses can set up automated safety measures that can react to detected anomalies. For instance, if an abnormal traffic pattern is identified. Security is the last barrier for defending ourselves from threats. Indeed, deploying extra security functionalities seamlessly across CNFs without service interruption, critical services will remain secure against emerging threats.

- Building Long-Term Relationships: Trust-centric capabilities like those provided by CNFs and user-centric designs help businesses build long-term relationships with customers by demonstrating a commitment to protecting their data and delivering consistent, high-quality service.

While there is still a lot of work to be done in this regard, the development of TFs in the form of CNFs in 6G networks offers businesses enhanced resilience, reliability, privacy, safety, and security. These capabilities ensure consistent service quality, protect sensitive data, and adapt to user needs in real time, making them crucial for customer trust. Eventually, services could quickly scale, improving response to disruptions, and secure data processing at the edge. By prioritizing these TFs, companies can differentiate themselves in competitive markets, offering tailored, high-quality services that attract and retain customers. Besides, companies may quickly introduce new offerings and tap into premium markets, thus driving increased income. At the same time, the modular architecture reduces operational costs through automated recovery, streamlined updates, and optimized resource usage. By leveraging these efficiencies, businesses can boost profitability while maintaining high service quality and customer satisfaction.

5 SAFE-6G Overall building blocks and Components

In this section, the main building blocks that constitute the SAFE-6G architecture are introduced, designed to enhance user-centric trustworthiness within the 6G ecosystem. This overview includes the User Intent Large Language Model, Chatbot, Cognitive Coordinator, XAI, Trust Functions, Edge Cloud Continuum, and MLOps/DataOps framework, each playing a pivotal role in enabling real-time, adaptive, and transparent trust management. These components collectively support a high level of customization, resilience, and trustworthiness within the SAFE-6G architecture, paving the way for a secure, distributed, and user-centric 6G network framework.

5.1 User intent LLM

5.1.1 Overview of the Chatbot in SAFE-6G.

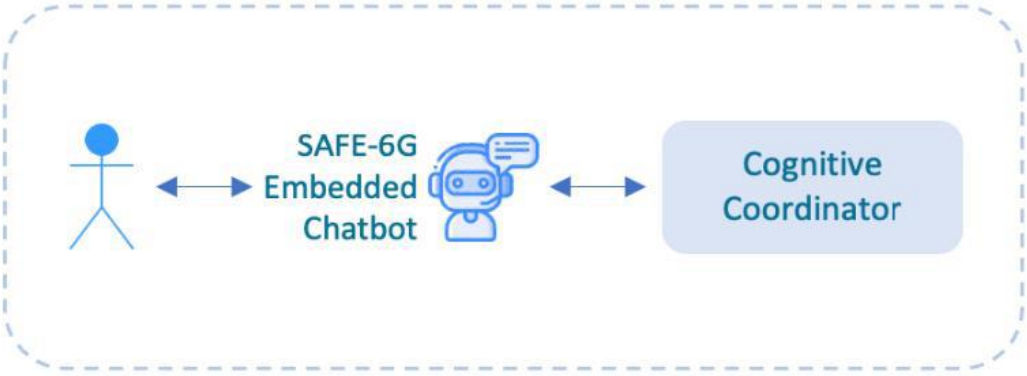

Figure 21: A high-level view of the chatbot.

The Chatbot component within the SAFE-6G framework is the first domain that contributes to the enhancement of the user interaction and ensuring trustworthiness in the User-centric Distributed 6G Core. More specifically, it will allow users to interact with the system to request the LoTw needed through the SAFE-6G Cognitive Coordinator, that extracts the LoTw across the five SAFE-6G user-centric functions: safety, security, privacy, resilience, and reliability. The procedure starts by prompting service-driven questions to the user. Then, this acts as an input and is processed by the NLP component that converts the raw text into structured data that is later an input to the Cognitive Coordinator Component. Additionally, this AI chatbot will be embedded within the immersive environment of the metaverse pilots, so the user can request the LoTw he needs. He will also receive XAI responses on the AI models that are going to be applied to achieve the LoTw.

5.1.2 User Interaction and Trustworthiness

The chatbot facilitates the user interaction through the intent classification. Intent classification identifies the primary goal of the user. These intents are mainly domain dependent. For example, a request can be in our domain of the SAFE-6G environment, privacy provision, robustness of security, and so on. The chatbot supports English language.

Examples in the Figure 22 below illustrate the dialogue between the user and the chatbot and intent classification processes described above.

Input	*Intent*
Upgrade my security	**Security**
Predict resource needs reliably	**Reliability**
Select best trusted paths for privacy and efficiency	**Privacy**
Ensure safe access to IT service management platforms	**Safety**
Create ontology of resilient functions	**Resilience**

Table 4: Examples of the users-chatbot interaction and intent classification

The queries can be in the form of unstructured natural language inputs. These unstructured user queries are then processed and parameterized by the Cognitive Coordination Component. Then the Cognitive Coordinator extracts other necessary details, which when combined with the intent, allows it to fully understand the user's request. Eventually, through this interaction the user's trust will be ensured. In case any necessary parameter is missing from the user input, the chatbot engages in conversation with the user until all the parameters are provided correctly (Figure 21, Figure 22). The chatbot responses can be in the form of text.

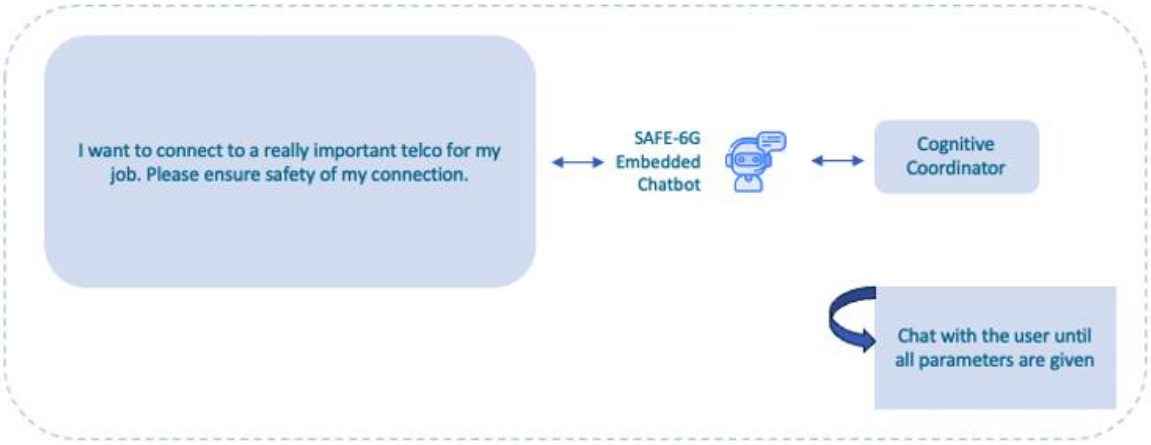

Figure 22: An example of how the user query is handled by the chatbot.

5.1.3 Integration with the Metaverse

The chatbot works as an interface between the IMM metaverse application and the Cognitive Coordinator and the users. Within the metaverse applications the user can interact with the cognitive coordinator of the SAFE-6G system and select the level of trust, but also to receive XAI on the AI/ML models that are going to be applied for achieving the requested level of trust. End users will be able to interact with the SAFE-6G chatbot, represented as a virtual chat window in Extender Reality (XR). The end user inputs will come from the vertical applications of the two metaverse use-case applications and be sent to the Safe-6G chatbot. The envisioned web APIs between the Chatbot and the metaverse applications will be presented in D2.3. Explainability (XAI) in AI Chatbot.

Achieving LoTw will require decisions to be taken by AI Agents defined in the MLOps layer. In an architecture where AI models must make critical decisions—whether about resource allocation, prioritization, or real-time responses—users need assurance that these decisions are not only correct but also transparent and explainable. Without this, the AI system risks losing the trust of its users, leading to underutilization or outright rejection. To maintain user confidence and ensure that the Cognitive Coordinator aligns with user requirements, the Cognitive Coordinator will access the XAI module to assess black-box decisions and selectively present these insights back through the AI Chatbot. The XAI module thus needs to integrate with Generative AI capabilities to interact with the user through the Coordinator chatbot and to trigger XAI techniques – cf section _XAI_(THA).

5.2 Cognitive Coordination

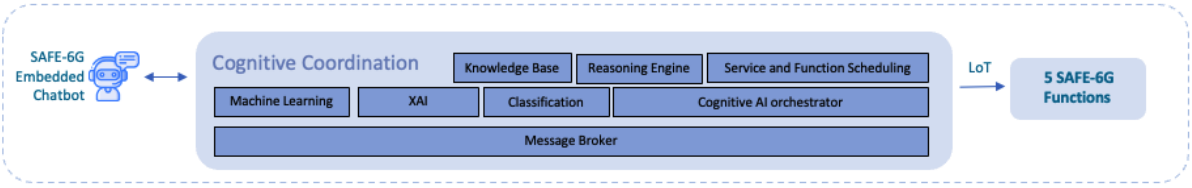

Figure 23: SAFE-6G High level view of Cognitive Coordinator

The SAFE-6G Cognitive Coordinator is an advanced intent-handling component that plays a key role in managing the trustworthiness of the SAFE-6G user centric cognitive framework. It interprets user-defined trust requirements conveyed via the AI Chatbot, calculates the desired LoTw and coordinates the system's transition to a trustworthy state. This is achieved by orchestrating the five SAFE-6G trust functions: safety, security, privacy, resilience, and reliability.

The Cognitive Coordination process involves managing and orchestrating the classification and reasoning activities to ensure optimal and achievable LoTw in the SAFE-6G framework. Initially, the regression component produces a non-calibrated level of trustworthiness (nLoT) by extracting features, processing data, and running an AI algorithm on the input. The Reasoning Engine then refines this initial nLoT,profiling each trust function and querying its knowledge base to integrate user info, system metrics, and understand the corresponding actions the TFs should perform to proceed with the deployments. Here is the point where any conflicts are identified and resolved. The outcome of the reasoning engine is a calibrated level of trustworthiness (cLoT) that corresponds to feasible deployment actions. This cLoT is communicated to the TFs via a message broker which in return performs the actions needed. There are five different key components, as depicted in Figure 23, in the cognitive coordinator: regression component, reasoning engine, cognitive AI orchestrator, SAFE-6G trust function, and the XAI component.

Following the role of each component and the process flow is described:

1. **Regression component:** The process begins with the regression component, which handles data received through the AI Chatbot. It uses an AI algorithm to quantify the received information from the chatbot and generate a nLoT. This nLoT acts as the initial trust metric, which will be further refined in the next stages.

2. **Reasoning engine:** The reasoning engine is composed of the Knowledge Base and the Conflict Resolution Manager and is tightly bound to them. It utilizes information about the users, resource constraints and deployment actions mapped to trust scores stored in the knowledge base to refines the initial nLoT to cLoT by inferencing the Conflict Resolution Manager to identify and resolve any conflicts. This way, it ensures that the proceeding actions are aligned closely to the user's expectations while maintaining them in a feasible spectrum.

3. The **cognitive AI orchestrator** is the central component that connects all subcomponents within the Cognitive Coordinator. Also, it includes a message broker which is responsible for the communication of the internal components of the cognitive coordinator as well as the external components such as the trust functions.

4. **Trust functions:** Each of the five trust functions' agent receives the calibrated trustworthiness level coefficient and performs any actions needed to achieve the dictated-by-CoCo trustworthiness level. These functions work independently but are coordinated through the system to collectively uphold user-defined trustworthiness standards.

5. **Explainable AI (XAI):** Throughout the process, XAI provides clear explanations of how decisions are made to achieve the desired trustworthiness. This transparency builds user confidence and understanding by clarifying the reasoning behind system actions.

Together, these components enable the SAFE-6G cognitive coordinator to interpret user intent, refine trustworthiness levels, and coordinate actions across the system, ensuring that the framework aligns with user-defined trust goals.

5.3 XAI

Black-box models, typically deep learning architectures or complex ensemble methods, are often highly accurate but lack the transparency that users require to fully trust their outputs. These models do not provide a clear, understandable explanation for how they reached a decision, especially when compared to simpler, rule-based systems or interpretable models like decision trees. As a result, even though black-box models often outperform other models in terms of predictive power, they are more challenging to deploy in settings such SAFE-6G that demand high trust and accountability.

To address this problem, the XAI component will analyse the results of the five SAFE-6G trustworthiness functions and will propose explainable results to the end-users. The XAI component proposes two kinds of approaches:

- ***Local explanations***: the goal here is to understand a particular value returned by one of the 5 functions. For instance: if a Level of Trust cannot be achieved in the current environment, what would be the minimal change in the network architecture required to achieve the target Lot?
- ***Global explanations***: the goal here is to understand the global behaviour of the model. For instance: on what basis/feature the model splits between predicting low performance and high performance of a given topology?

The component is also in close relation with the Cognitive Coordinator module and with the chatbot to provide explanation on its decision-making process to final users.

Here are some of the AI techniques that may be implemented to support the Cognitive Coordinator's role in enhancing explainability:

1. **Model-Agnostic Methods**: Techniques like SHAP (SHapley Additive exPlanations) provide interpretable approximations for black-box models. These methods allow the Cognitive Coordinator to generate human-understandable explanations even when the underlying model is complex. They do so by creating a simplified model that mimics the black-box's behaviour within a specific decision-making context.
2. **Feature Attribution**: Techniques like ALE (Accumulated Local Effects) that attribute a model's decision to particular features or inputs can help users understand which factors most heavily influenced an outcome. The Cognitive Coordinator can present this information to users in the form of ranked features or contributing variables, enhancing the system's transparency. In healthcare applications typically, these helps knowing which patient symptoms led to a given diagnosis by the AI. Similarly, in SAFE-6G, these features could allow the Cognitive Coordinator report – via the Chatbot – that the resilience of the network can be mostly attributed to a given network property.
3. **Counterfactual Explanations**: These explanations provide insights by illustrating how a decision would change if the input data were slightly different. By showing users what conditions might have led to a different outcome, the Cognitive Coordinator enables a deeper understanding of the AI system's decision logic. For instance, if a given cLoT is not achievable with current network configuration, the Cognitive Coordinator could rely on this module to explain what would be missing in the network configuration to achieve it.
4. **Rules Extraction from Neural Networks**: For deep learning models, particularly neural networks, rules extraction techniques attempt to derive logical rules or simplified decision trees that approximate the network's decision-making. Though these rules may not perfectly capture the black-box model's internal workings, they can provide a reasonable degree of interpretability, making them useful in enhancing transparency.

Implementing these techniques within SAFE-6G networks presents numerous challenges. Key open challenges for this component include:

1. **User adaptability**. Explainability involves more than merely providing a glimpse into the inner workings of AI models; it requires delivering explanations tailored to the user's expertise, objectives, and needs. For example, while developers and regulators might need highly detailed, mathematically intricate explanations, non-expert users may benefit from simpler, more intuitive descriptions.
2. **Common ground knowledge**. It is crucial not only to identify the features used by the model for its predictions but also to make these features comprehensible to users. Which network features will be utilized by predictive models and how to make these features understandable to users remain unresolved challenges.

3. **Interpreting the results or scores generated by XAI techniques**. This could involve summarizing these outcomes through visual diagrams or providing textual interpretations to make the insights more accessible and understandable.
4. **Computational efficiency is also a concern**. Many XAI techniques require access to large datasets and involve complex computations, which can delay the delivery of explanations. This delay could disrupt real-time interactions, such as chat-based discussions. While the MLOps component may help by triggering offline pre-computations as a preprocessing, it remains uncertain how frequently this will be feasible or optimal for maintaining a seamless user experience.

The XAI component requires interacting with the Cognitive Orchestrator and propose results that can be used by the Chatbot to interact with users. Thus, it needs to integrate with some Generative AI capabilities, which will be used both to understand user explanation requests and to provide user-adapted interpretations of XAI technique results. In addition, triggering and configuring XAI techniques will require close interactions with the MLOps framework, to provide information on how models were trained, and to schedule some XAI jobs.

The integration of the XAI component within the architecture ensures that explainability is not an afterthought, but a core feature of the system's decision-making framework, and that it brings meaningful answers to users through the Chatbot.

5.4 Trust Functions

The TFs in SAFE-6G are designed to ensure that the network maintains its overall trustworthiness by dynamically managing and optimizing several key dimensions: security, privacy, resilience, reliability, and safety. These functions are deeply integrated with the network's cognitive systems and are executed through local AI agents, making them adaptable to real-time conditions and focused on the overall network's trustworthiness.

At the core of the SAFE-6G architecture is the concept of cLoT, which represents the level of trust the system maintains at any given time. The TFs are directly responsible for calculating, managing, and enhancing this trust level by considering user needs but primarily ensuring that the network meets the desired LoTw in different operational contexts. More detailed attention to balancing individual user-specific needs with overall network-level trust is crucial as network personalization increases.

5.4.1 The SAFE-6G Trust Functions

The SAFE-6G TFs are responsible for ensuring the overall trustworthiness of the network by continuously assessing and calculating the cLoT. These functions operate at a higher level, focusing on providing a trust score that reflects the system's current state of trustworthiness. They work dynamically, leveraging AI-driven mechanisms to adapt to changes in the network's environment, threats, and operational requirements.

The TFs can encompass several critical areas such as security, privacy, resilience, reliability, and safety, but their main purpose is to provide an aggregate trust score, rather than focus on the specific trust needs of individual users. While the TFs may incorporate some user preferences, they are primarily

designed to manage trust at the network level. Further refinement is needed to explore how the user preferences are dynamically integrated into the TFs without compromising the overall trust level depending on their individual baseline scores respectively.

The safety aspect evaluates the network's ability to ensure that users and systems are not put at risk. It provides a safety trust score based on how well the network prevents unsafe operations, especially in critical environments like autonomous systems.

The security aspect of the TFs evaluates the current state of the network's protection mechanisms, such as encryption and access control, against potential threats. It calculates a security trust score based on how well the network is protected against unauthorized access and cyberattacks.

Privacy is another key component of the TFs. This aspect evaluates how well the network complies with privacy regulations and user preferences. The TFs monitor data handling and storage to provide a trust score that reflects the privacy safeguards in place. A transparency mechanism for informing users of privacy handling and how their data is safeguarded could improve user confidence and compliance with privacy regulations.

The resilience function evaluates the network's ability to withstand and recover from disruptions. This function provides a resilience trust score, which reflects the network's capacity to adapt and continue operations in the face of challenges such as hardware failures or cyberattacks.

The reliability function focuses on ensuring the reliability of the service during its runtime. It uses data from the different planes to feed ML/AI methods and perform inspection for abnormalities or malicious actions. It also generates alerts when a malicious action, abnormality or QoS violation is detected.

To account for varying operational contexts, dynamic weighting of the trust dimensions (such as resilience, reliability, security, privacy, safety) will be considered to enhance the flexibility of trust calculations performed.

5.4.2 Dynamic Trust Score Calculation

The TFs in SAFE-6G work through a combination of real-time data monitoring and AI-driven decision-making to calculate the achievable cLoT. The trust score dynamically adjusts based on the current state of the network and external factors, providing a continuous assessment of the network's overall trustworthiness. This score is essential for the network's cognitive architecture, ensuring that trust levels are maintained across various dimensions, including security, privacy, resilience, reliability, and safety. To make this process more transparent, explicit methodologies for calculating cLoT, including AI models, data sources, and algorithms, should be defined to ensure trust and fairness.

While these functions ensure the system's trustworthiness at a high level, they remain flexible and adaptable, allowing the network to respond to new threats or changing conditions in real-time. Although the system considers user needs in the trust calculations, the TFs are designed to ensure that the network's overall trustworthiness remains optimized.

5.4.3 Role of AI in Trust Functions

The integration of AI plays a crucial role in the SAFE-6G trust functions. Each TF is supported by a local AI agent responsible for mapping the desired trustworthiness score (e.g., 15%, 50%, 80%, etc.) to an appropriate trustworthiness level (e.g., low, medium, high, etc.). Once the trustworthiness level is determined, the function communicates with the cognitive coordinator to deploy the corresponding network app which is responsible for achieving this trustworthiness level. The local AI agents continuously monitor the network's state and adjust trust scores based on real-time analysis. However, the TFs do not directly modify the LoTw themselves; they communicate with the Cognitive Coordinator to make necessary adjustments. These local AI agents gather data on performance, security events and user behavior, enabling the TFs to respond dynamically to changing conditions.

AI agents within the TFs are responsible for:

- **Monitoring**: Continuously collecting data on network health, security risks and operational performance.
- **Decision-Making:** Using this data, AI agents assess the current cLoT and determine whether the system is operating within acceptable trust parameters.
- **Action**: When trust is compromised, AI agents take corrective actions. This could involve tightening security protocols, reallocating resources or mitigating emerging risks.

Additionally, AI plays a critical role in the network applications themselves, such as in anomaly detection models. These models continuously analyze real-time data to identify unusual patterns or behaviors that may indicate security threats or performance issues. When such anomalies are detected, the network app can automatically trigger protective actions, such as isolating suspect traffic or reconfiguring security protocols.

To ensure transparency and accountability in AI decision-making, providing explainability and an AI audit trail will be essential, especially in cases were corrective actions impact users or critical network operations.

By embedding AI into the TFs, the SAFE-6G system can dynamically adapt to fluctuating conditions and maintain optimal trust levels. This approach ensures that the system remains resilient and secure, responding automatically to potential trust breaches or threats.

5.4.4 Lifecycle Management of Trust

The TFs manage trust throughout the entire lifecycle of network services, from deployment to operation and decommissioning. This lifecycle management ensures that trustworthiness is maintained as the network evolves and changes.

When new services or devices are added to the network, the TFs evaluate their trustworthiness. AI agents assess whether the new components meet predefined trust criteria before they are integrated into the system.

During network operation, the TFs continuously monitor the system's trust scores, responding to new threats, changing user demands and evolving network conditions. AI-driven trust management ensures that the system adjusts in real time to maintain trustworthiness.

When services or devices are removed from the network, the TFs ensure that trust is maintained during the decommissioning process. This includes secure data handling, ensuring that no residual data compromises network security or privacy.

More specific strategies for managing trust during scaling, load balancing, or handling the integration of new devices would enhance the robustness of lifecycle management.

By managing trust across the service lifecycle, the TFs ensure that the network remains secure, resilient, and reliable throughout its operation. The use of AI-driven trust management allows the system to adapt and respond dynamically to the needs of the network and its users.

5.5 Evolution of Core Network to Distributed Ecosystem

The Core is evolving towards fully distributed system (Figure 24) where network functions are deployed on customer premises or in private/public clouds. The core can assign different network functions for different users through slices or subnets.

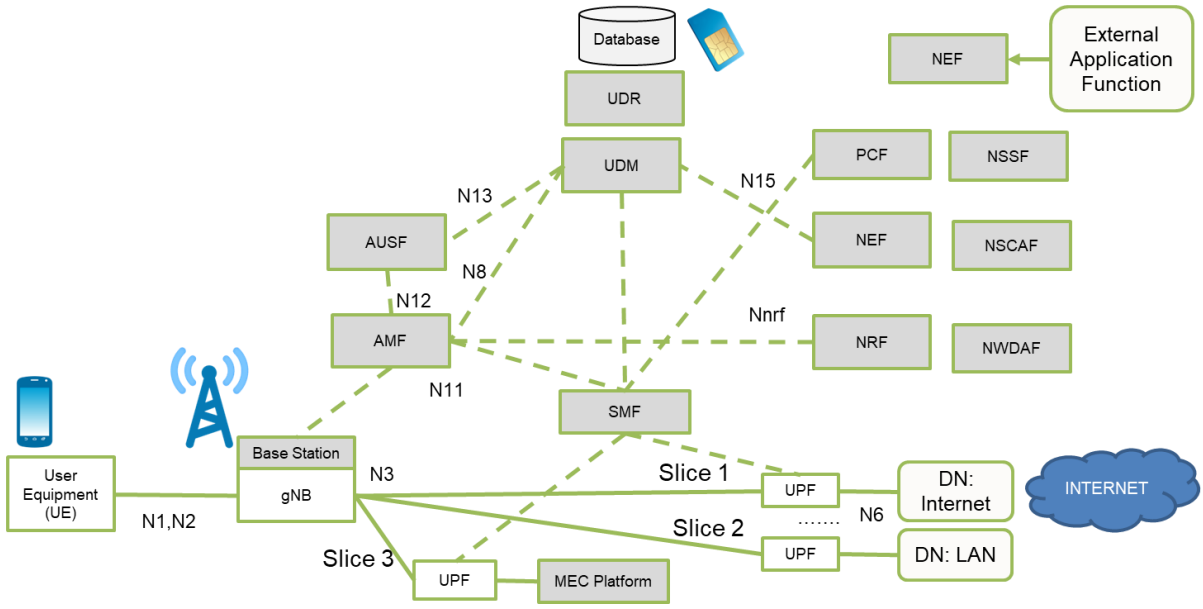

Figure 24: Architecture of a 5G network.

The packet core is designed to meet the needs of modern private mobile networks. At the heart of this system is its modular and open architecture, enabling seamless integration and customization across a range of use cases. The Packet Core is built using microservices, with each network function—such as the AMF (Access and Mobility Management Function), UPF (User Plane Function), and SMF (Session Management Function)—operating independently in line with the 5G architecture. This modular approach not only increases flexibility but also allows for the efficient scaling of resources, making it easier to meet the demands of diverse network slices based on user specific services.

The platform offers a comprehensive set of open APIs that enable easy integration with third-party applications, orchestration tools, and process automation systems. These open interfaces allow external systems to interact with the network core, creating opportunities for real-time control, monitoring, and customization through Slicing mechanism. This makes the platform ideal for experimentation in slicing, automation, and network management. By leveraging open APIs, researchers and developers can quickly adapt and test network functions, contributing to faster innovation and optimized network performance.

Network Slicing feature enables the creation and delivery of multiple virtual networks from a single physical network infrastructure. These virtual networks, or slices, can be customized with distinct operational characteristics to suit various use cases and requirements and different trustworthiness levels. Each network slice can differ in terms of maximum user security, trustworthiness, capacity, data throughput, and access to specialized features, providing a flexible platform for tailored services. One of the most powerful aspects is its ability to dynamically create and manage network slices. This allows network operators to allocate slices for specific roaming operators or organizations, ensuring that roaming users have secure, dedicated access to network resources. By segregating network traffic and resources, operators can offer enhanced performance and reliability to roaming partners without affecting the overall network's operation.

Network Slicing (Figure 25) is achieved by sharing the RAN while dedicating transport layers per slice. As depicted in previous figure the core network, at least the UPF is dedicated per slice, serving as a critical integration point for external data networks and enabling seamless data exchange for roaming operators. As needed, other virtual network functions can also be assigned to specific slices, ensuring granular control and optimization of resources.

Slices are organized per user or group of users per their communication needs. This makes QoS management significantly easier to manage and also Security and Service Level Agreements.

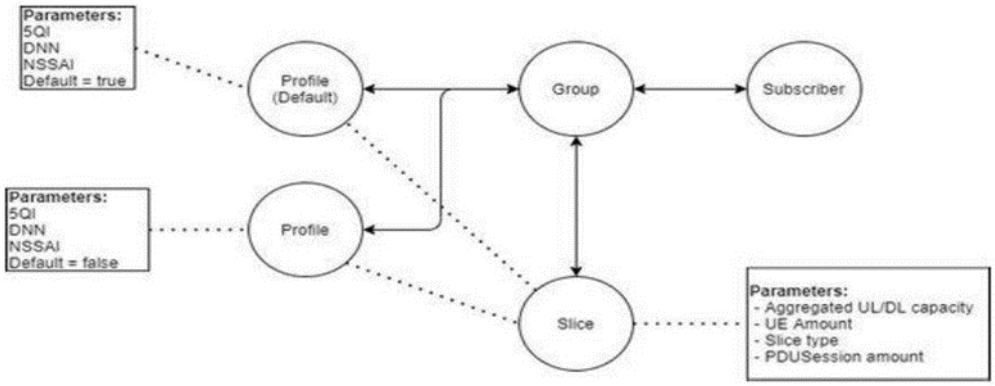

Figure 25: Network slicing model

5.6 Edge cloud Continuum

The Edge Cloud Continuum (referred to it also such as Cloud-Edge-IoT continuum or computing continuum) is a paradigm that represents the convergence between the heterogeneous capabilities

available at different computing layers (Cloud, Edge, IoT). This convergence allows the operation of workloads over the distributed resources available at the different computing layers, of which involved services and applications can be allocated depending on their specific requirements (e.g., latency, availability, acceleration capabilities, etc.).

The EC is fostering this paradigm, channeling the effort of several research and innovation projects through initiatives such as EUCEI. This initiative has been working on a common taxonomy, reference architecture and set of expected functionalities for the continuum. The SAFE-6G Consortium firmly believes that the next evolution of cellular technology must embrace this paradigm to foster the implementation and management of greener, more advanced, and customized network slices, services, and applications. In this paradigm, computing elements (referred to as **Infrastructure Elements, IEs**) can be grouped into **domains**, which can represent aspects such as ownership, location, or layer (edge, fog, cloud).

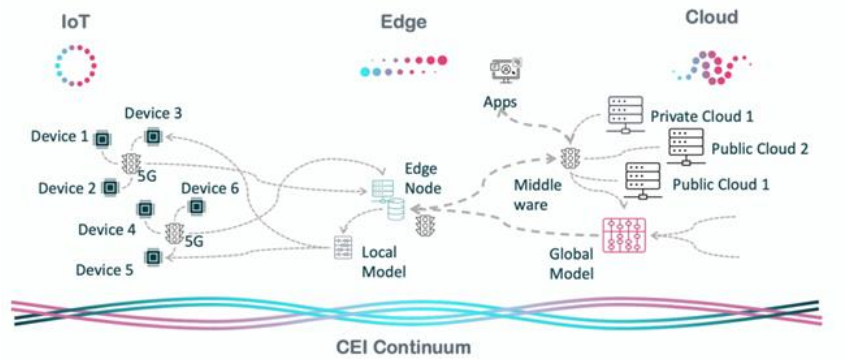

Figure 26: Computing continuum perspective.

While the continuum can bring several advantages, it also comes with a layer of complexity. Not only in terms of having a plethora of heterogeneous devices with different computing capabilities, but also potentially distributed in several locations and owned by different stakeholders. Therefore, **interoperability** challenges need to be carefully identified and addressed so that the ecosystem remains manageable, for instance:

- **Computing**: The continuum may host constraint IEs at IoT and edge layers as well as powerful ones at Cloud, in terms of processing power, memory and AI acceleration.
- **Virtualization**: While Kubernetes is becoming the *de facto* container orchestration platform and is taking over the role of Virtualized Infrastructure Manager in MANO, there are several distributions and, additionally, some devices (mostly constrained ones) might leverage only Docker or containerd. Besides, technologies for deploying services may also vary.
- **Networking and service discoverability**: Layer 3, 4 & 7 technologies (e.g., VPNs, CNIs, and service mesh) may be managed and exposed differently in each domain.
- **AI**: Some IEs might have specific platforms to train and serve AI models, while others may have other similar tools or lack them. Additionally, models may only run on specific ones.

Management: considering all the previous (plus security and data interoperability) challenges, orchestrating resources and services requires specific, adaptable solutions.

To address these interoperability challenges, one approach is to consider a **Meta-OS**. In this context, a Meta-OS can be described as a dedicated set of services to be installed on the computing resources of a continuum, on top of their respective operating systems, delivering a concrete set of functionalities related to the management of a distributed edge-cloud computing continuum. Depending on the specific Meta-OS implementation considered, features related to service orchestration, resource and network management, data management, security and privacy, trust and reputation, and monitoring and observability can be enabled through the continuum, supported by AI. The Meta-OS components are in charge of abstracting the technologies involved (i.e., data models, API endpoints) and simplifying their management through unified interfaces, accessible via secured Open API or management portal.

Aligning with the current Cloud Native trend in the telco realm (based on pillars like microservices, containers, Kubernetes, DevOps, and CI/CD), the Meta-OS should also perform based on those principles. Among the features and functionalities that can be introduced by a Meta-OS, the ones described in the following sub-sections should be enabled to ease the deployment and execution of the remaining SAFE-6G system. It should be highlighted that the Meta-OS is not the only element of the SAFE-6G ecosystem that addresses interoperability challenges. Other technologies, such as CAPIF, also contribute to them, in this case by unifying the exposure and authentication of the APIs of the lower planes of the framework (i.e., continuum, network core and use cases' APIs).

5.6.1 Federation and Orchestration

Federation is a feature that facilitates the discovery and sharing of resources and services as well as the exchange of data among domains belonging to a given continuum. It consists of a set of services installed on each domain which essentially unifies the network, data, and service fabrics of the continuum into a scalable, cohesive system. This feature is essential for the rest of the Meta-OS functionalities to operate seamlessly. One approach is to register a domain into an administrative or entrypoint domain, which then synchronizes its relevant data with the rest of domains, so they are also aware of the existence of it.

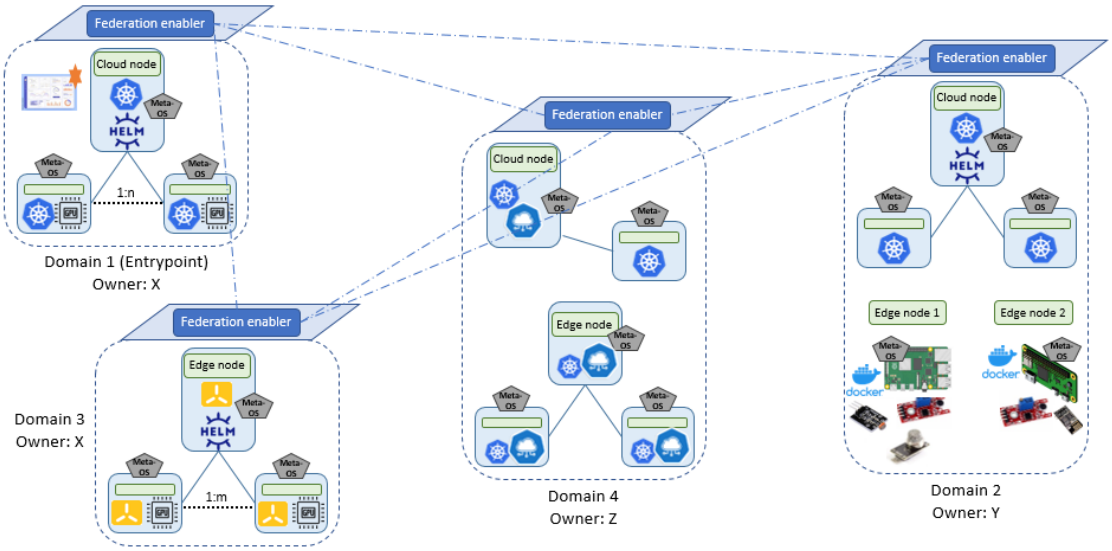

Figure 27: Federation of computing continuum domains.

The **orchestration** module leverages the continuum established by the federation to manage and orchestrate the lifecycle of services and applications throughout the different domains of the continuum. When a third-party system (for instance, the SAFE-6G cognitive coordinator, a 5G/6G core, end user applications, etc.) is to be deployed, this module decides the optimal IEs on which deploy the involved services regardless of the domain they belong, or the virtualization technologies considered (yet considering its computing capabilities and the needs of the service/s to allocate). It is also in charge of upgrading, downgrading, deleting, or rescheduling them. To realize it, SAFE-6G follows the two-level orchestration model fostered by EUCEI, which consists of two main services:

- High-Level Orchestrator (HLO, one per domain). It receives a deployment request and decides the optimal IE to instantiate the services (depending on requirements, real-time information, forecasts, etc.), supported by AI. When the IEs that are involved are selected, it communicates with the following component to deploy them – for which federation is key. It can also forward requests for upgrading and deleting services.
- Low-Level Orchestrator (LLO, one per IE). When it receives a request from an HLO, it interfaces with the underlying virtualization technologies available (Docker, K8s, K3s, KubeEdge, Helm, Juju, OpenStack, etc.) to deploy, upgrade or delete a service. Notice that the HLO does not care about the virtualization technology of the IEs to make its decision.

In SAFE-6G, the coordination module will be leveraged to deploy the rest of the SAFE-6G system and will be interfaced by the latter once a request is received via the chatbot, after being processed by the cognitive coordination of the trustworthiness framework. Once a request is received (via Open CAPIF interface), the balancer of the entrypoint domain selects the HLO of the continuum that will make the required service allocation decisions, which in turn will communicate with the relevant LLO/s to deploy them (which, as mentioned, may or may not belong to its domain).

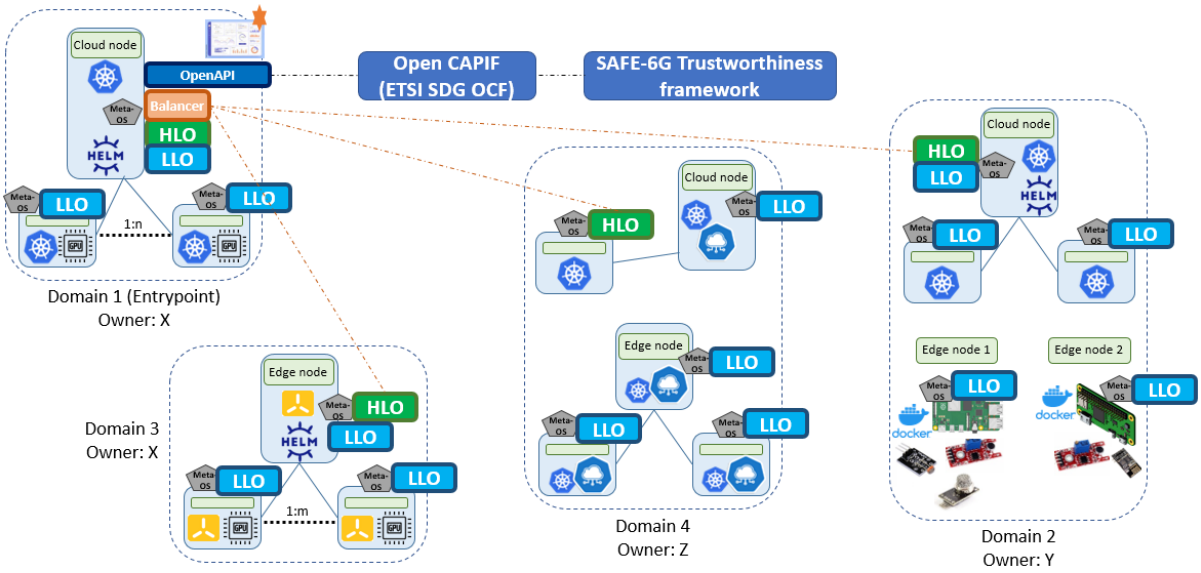

Figure 28: Orchestration of services in SAFE-6G's computing continuum.

5.6.2 DATA MANAGEMENT

While this module can integrate several data-related features, the crucial component is the **Data fabric**. It allows any data consumer of the continuum to request and receive data from other components, without the need to interact directly with the domains or the producers of such data (via push or pull queries, through multiple mechanisms). Data owners can publish data into the fabric directly, if it is already aligned with the proper data model and protocols from its source, or it can be processed to serve it properly. Rights over the managed data can be also specified. Federation is key so that the different instances of the data fabric deployed in the different domains can communicate among them, so data is available across all the continuum without replicating it.

Apart from other modules of the Meta-OS (like the HLO), a data consumer can be other elements of the SAFE-6G trustworthiness framework that require real-time data from the continuum (like MLOps or the trust functions), or end user applications. While access control mechanisms might be natively available in the Meta-OS, SAFE-6G will consider Open CAPIF as the primary mechanisms for publishing and making data available for the remaining building blocks of the architecture.

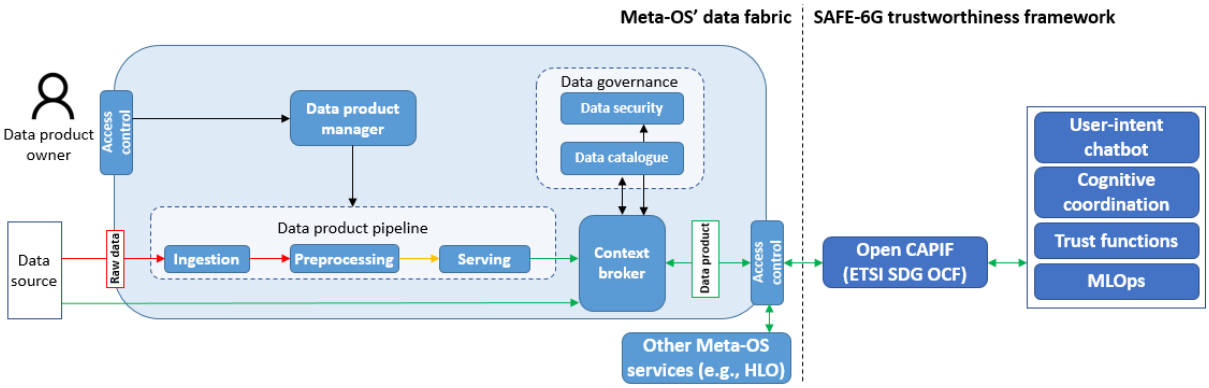

Figure 29: Data fabric components and interaction with the rest of SAFE-6G's building blocks.

5.6.3 MONITORING

The monitoring module has the role of collecting all the relevant metrics so that the orchestration module can make well-informed decisions (e.g., CPU, RAM, GPU resources available, latency figures, power consumption, etc.). Other autonomous capabilities of the Meta-OS, like self-diagnosis, self-recovery, or self-scaling, depend on having metrics properly gathered (via **monitoring agents**) and published into the data fabric.

While Meta-OS gathers metrics that can be of interest for the TFs, others might be required so the functions and overall, the SAFE-6G trustworthiness framework can provide their expected features. The necessary monitoring agents or probes must be then developed, deployed and the gathered metrics properly ingested into the data fabric, so they can be seamlessly accessed by the rest of SAFE-6G components.

5.7 MLOPS

The MLOps paradigm aims to simplify the development of machine learning models, as well as their deployment and maintenance. This is achieved through the creation of automated processes, which make the development of AI models more reliable and productive. A well-configured MLOps environment provides security for the entire team, from those responsible for creating the models (data Scientist, ML engineer…) to those responsible for deploying and maintaining the infrastructure, reducing production time at every stage.

The MLOps Framework proposed in SAFE-6G, manages the entire lifecycle of distributed AI/ML models, including their creation and orchestration. The framework is infrastructure-agnostic, allowing for in-network training of models. Supported by an XAI module, it provides users with understandable explanations about the trust level offered by the cognitive coordinator layer. The training datasets are sourced from monitoring modules that aggregate data from the network continuum and are supplemented by a DT component when real data is insufficient. To protect privacy, a Differential Privacy (DP) module adds noise to the datasets without compromising model accuracy.

The SAFE-6G's MLOps framework can be seen as an abstraction layer on top of cloud-native architectures managed by Kubernetes, which allows developers of AI models to work on a simple platform by abstracting them from everything that happens internally: execution of Kubernetes runs, deploying services to serve the models, computational resources used for training and inference, etc. The MLOps module will run on local Kubernetes clusters or in the cloud, leveraging all available resources to accelerate training and deployment.

The MLOps framework will consist of a gathering of several open-source projects that solve by themselves different stages of ML lifecycle. The cornerstone of the framework will be Kubeflow, an open-source platform based on Docker containers and Kubernetes for building ML workflows. Another series of modules will be developed around this platform to provide the framework with the desired functionality, such as Minio, Tensorflow Serving or Torch Serve and so on. The different modules that compose the whole MLOps framework are described in following sub-sections.

5.7.1 PIPELINE DEVELOPMENT

Constructing an IA model is not a one-way process, but an iterative workflow in which different configurations must be tested until the final solution is found. These iterative processes occur at different stages of the lifecycle of a machine learning model: for example, when choosing the training and evaluation data, when applying preprocessing over them, or when designing the model's architecture or the hyperparameters with which to train it. Therefore, any development of an IA model requires a test platform on which to conduct the experiments.

The pipeline development module is a workplace with several tools to help the user/customer doing data exploration and model development, whereas they are defining the different stages of and MLOps pipeline. It is based on Kubeflow Notebooks, a sub-module of Kubeflow which use internally JupyterLab as development platform. JupyterLab is a web-based application that offers an interactive working environment through Jupyter Notebooks. The main advantage of working with Notebooks in

Kubeflow is that each user can create their own working environment using docker images, unlimited development scope.

This module provides the user with a test environment in which to carry out experiments. Once a final solution is available, the next step will be to build a production pipeline that will be used in the next module as an automatic process to build an AI model from scratch. The steps to build the production pipeline will be discussed in the following section.

By using the pipeline development tool, these could be an example of the steps that the user would have to carry out:

I. Prepare a docker image with all the necessary packages to be able to conduct the development of the AI model. This image will be used as the development environment in the Jupyter Notebook.
II. Start with the model development tasks in the notebook itself, such as preparing the data, choosing the model architecture, start with the first's trainings, etc.
III. Configure the pipeline itself that will be orchestrated later in the pipeline orchestration platform (discussed in the following section).

Additionally, with the multi-user isolation function Kubeflow lets various users to log into the platform and allocate separated workspaces, one per user.

5.7.2 Pipeline Orchestration Platform

In the previous section it can be showed how a user could use the pipeline development module as a test environment to build an AI model. The first trainings could be undertaken there but not before having done some research on data pre-processing and different model architectures. All these steps should be contained in code, and it is time to sort them out in a ML pipeline. A machine learning pipeline is defined as the individual steps that must be completed in order to build an AI/ML model from the ground up.

The Pipeline Orchestration Platform (POP) is based on Kubeflow pipelines, which allows the user to run, schedule and monitor AI/ML pipelines in the form of Direct Acyclic Graphs (DAGs) programmatically and sequentially. At runtime, each step in the pipeline corresponds to a single docker container execution which has associated inputs and outputs (denoted as artifacts). When the user executes a pipeline, the POP launches several Kubernetes Pods per each component in the pipeline, which are in charge of running the applications inside the containers developed by users.

To set up the pipeline, several components are required. The first of these is a docker image that contains everything necessary for the creation of the AI model, normally the same image has been used in the previous stage. It is even possible to use a specific image for each of the stages. The second corresponds to the code, Python scripts that are going to be executed in each of the stages of the pipeline (training scripts, evaluation, inference...), the whole code must be available in the docker image. Finally, you will have to make use of the Kubeflow pipeline SDK, a python library whose purpose is to provide an easy and intuitive way for the user to generate the pipeline. This library allows us to indicate the docker image that corresponds to each stage and the associated code to be executed.

The output of this stage will be a yaml file, which will have to be uploaded to the Kubeflow GUI in order to be able to run it in the future.

In summary, the POP module is in charge of orchestrating the pipeline that will be used to generate the AI models. The output of the pipeline will correspond to the trained model, usually a file with the architecture weights. This file must be stored to be able to serve the model and make inference on it. These functionalities will be provided by the following modules: 5.7.3. Model storage and 5.7.4. Model serving.

5.7.3 Model storage

Once the ML models are developed, trained, and properly evaluated, because of the pipeline execution, the next step is to store both the artifacts generated by the pipeline (architecture model weights) and the inference script necessary to apply the inference stage. The model storage module is a model registry based on MinIo for storing all the production-ready models and versions.

The storage module will be accessible by both the POP and the serving module. On the POP side it will be used to store files, while the serving module will use it in read mode.

5.7.4 Model Serving and Inference

The training of the models is not the last step in the chain, but it is necessary to encapsulate them to be consumed by the rest of the components of the SAFE-6G framework. There exist a couple of options for serving purposes which are part of the Model Serving high-level module which are based on TFS and TorchServe, one for the pipelines developed with TFX and TorchX, respectively. These microservices serve the desired model version/s which are stored at the model storage module.

The model serving module will expose model version/s through and API and the SAFE-6G components will be able to make inference to get model's predictions.

5.7.5 Differential Privacy

In the SAFE-6G project, a DP module is implemented to protect sensitive data collected from any node in the continuum. DP preserves privacy by adding random noise to data during model training, ensuring individual records remain anonymous while still enabling accurate analysis. There are two primary DP approaches: Central Differential Privacy (CDP), where a trusted server processes raw data and adds noise to query responses, and Local Differential Privacy (LDP), where noise is added directly at the data source, providing stronger privacy but at a cost to model accuracy. SAFE-6G focuses on applying LDP in machine learning models, aiming to limit accuracy loss to 4-5%, with a worst-case scenario of no more than 10% loss, ensuring privacy without severely compromising performance.

5.7.6 XAI components

The XAI component described in Section 5.3 interacts with both the Cognitive Coordinator and the MLOps framework. As the Cognitive Coordinator will trigger requests for explanation, the XAI component will need to retrieve information about models used by the AI Agents, their input features and how they produced their outputs. The MLOps Framework being the one responsible for producing and hosting these models, it is in charge of exposing this information to the XAI module.

5.8 DATA OPS

In addition, the XAI component will need to trigger some computations to provide interpretations. XAI techniques are in fact resources demanding operations, sometimes as much as a ML training phase. The MLOps framework being already equipped to orchestrate such jobs, the XAI component will use it to trigger XAI computations and retrieve results when they are ready for further interpretation.

DataOps is generally understood as a discipline that leverages a set of data-related practices, processes, and technologies to support an organization's data needs. In SAFE-6G, data is essential for the normal operation of the designed framework, from the training of the models involved (e.g., of the trust function, the chatbot, the explainability capabilities…) to the access to real-time data for their normal operation (e.g., of the Meta-OS, inference processes of the ML functions, etc.). This module offers data processed automatically so it can be leveraged by the different components of the framework. This data can be of two types: real and simulated. The first one will be collected from the planes (services, 5G system openness and edge-cloud continuum), whereas the second will be produced by the DT.

Figure 30: Components of the DataOps module.

5.8.1 MONITORED DATA

The monitored data block is responsible for gathering all relevant data, ensuring their persistency when required for the training of the ML models of the framework (or for feeding the DT). Monitored data should follow existing best practices of Cloud Native paradigm to foster its usability, for which Prometheus data model will be considered as it is the current *de facto* standard. Relevant metrics are:

- Metrics from the applications/services plane (with dedicated metrics exporters),
- Metrics from the 5G Openness plane (via NEF),
- Metrics from the edge-cloud continuum (via Data fabric).

This module can be designed and implemented considering different models and technologies, communicating with CAPIF as the main exposure mechanism of the planes. In the framework of this project, different alternatives such as SQL, NoSQL, time-series databases (such as InfluxDB or Thanos) and object storage systems will be considered for hosting the relevant data for training the ML models, gathered from all the planes. The selection depends on the data managed (e.g., tabular, time-series, voice, text, etc.). To homogenize the collection of metrics from services, Prometheus will be integrated as part of the overall solution. Since further preparation of the collected data can be dependent on the particularities of the consumers (e.g., models of the trust functions), any preprocessing needed by them will be managed within the MLOps module.

5.8.2 DIGITAL TWIN

In the context of the SAFE-6G user-centric cognitive framework, Digital Twin is a key component that enables the exposure and interaction of 5G network capabilities. SAFE-6G tries to emphasize adaptability, user experience and intelligent management, all of which are supported by this programmable network environment.

Within the SAFE-6G, there is a strong focus on dynamic, user-centric service customization. The DT's Exposure Layer, which simulates NEF APIs, aligns with this goal by providing an interface for third-party applications to access the 5G network's services. Through APIs for monitoring, analytics, and policy control, developers can interact with network elements that enhance the SAFE-6G framework's adaptive capabilities. This layer allows applications to respond to real-time changes, enabling customization of QoS and delivering services tailored to specific user needs and contexts.

Also, the DT will provide a virtual space where network behavior can be observed and tested under various conditions. This layer is crucial for SAFE-6G's cognitive functionality because it allows the framework to experiment with user-specific scenarios and generate data that can be analyzed for insights. For instance, by simulating mobile user behaviors and different traffic patterns, SAFE-6G can better understand user experience factors and dynamically adjust network resources to optimize service delivery.

By providing SAFE-6G with a flexible platform for continuous learning and adaptation. The framework's ability to make intelligent adjustments based on real-time network feedback is supported by the tool's capacity to simulate events and provide data on user interactions. SAFE-6G system can analyze these interactions to refine its cognitive mechanisms, improving network decision-making processes. For example, by simulating QoS degradations or location-based service adjustments, SAFE-6G can evaluate and enhance its response strategies, thus supporting a more resilient, user-centered network ecosystem.

6 CONCLUSION

This deliverable represents a pivotal milestone on the SAFE-6G project by introducing the overall reference architecture, a step toward achieving a user-centric 6G system rooted in native trustworthiness. Moving beyond a traditional focus on security, a comprehensive trust framework is introduced that encompasses safety, security, privacy, resilience, and reliability. Throughout the edge-cloud continuum, the SAFE-6G architecture prioritizes the cloud-native paradigm, guaranteeing a smooth integration and support for distributed AI/ML functionalities.

This architecture document incorporates six distinct perspectives—high-level, functional, process, data, deployment, and business views—each contributing to a comprehensive, multidimensional understanding of the system's goals and structure. This modular breakdown highlights the roles of key elements such as the user-intent chatbot, cognitive coordinator, trust functions, MLOps framework, edge-cloud continuum, and the 6G core, all essential to achieving the SAFE-6G trust framework.

The SAFE-6G's reference architecture is further enhanced by an AI/ML-assisted cognitive coordination component, which serves as an intent-handling function that interprets user trust intents across five trustworthiness dimensions. This component, together with the AI Chatbot enables users to request specific LoTw and receive transparency on the AI models engaged in delivering these trust levels. This alignment with user intent, coupled with XAI features, ensures that SAFE-6G meets the needs of a user-centric and intent-driven 6G environment, empowering users with control over their trust preferences and fostering a trustworthy and adaptable network ecosystem.

7 REFERENCES

[1] S. Angelov, P. Grefen y D. Greefhorst, «A classification of software reference architectures: Analyzing their sucess and effectiveness,» 2009.

[2] I. 42010, «ISO/IEC/IEEE 42010 - Software, systems and enterprise — Architecture description, Online: https://www.iso.org/standard/74393.html,» 2022.

[3] P. Kruchten, «Architectural Blueprints - The “4+1” View Model of Software Architecture,» IEEE Software 12 (6), pp. 42-50, 1995.

[4] «ISO/IEC 27001,» [En línea]. Available: https://www.iso.org/standard/27001.

[5] «33.501, 3GPP TS,» [En línea]. Available: https://www.3gpp.org/dynareport/33501.htm.

[6] «NIS2 Directive,» [En línea]. Available: https://www.nis-2-directive.com/.

[7] «EU 5G Cybersecurity Toolbox,» [En línea]. Available: https://digital-strategy.ec.europa.eu/en/library/eu-toolbox-5g-security.

[8] «ENISA 5G Toolkit,» [En línea]. Available: https://www.enisa.europa.eu/news/enisa-news/5g.

[9] ISO/IEC TS 5723:2022, "Trustworthiness - Vocabulary," International Organization for Standardization, 2022.

[10] NIST, "Framework for Cyber-Physical Systems: Volume 1, Overview," National Institute of Standards and Technology, Special Publication 1500-201, 2016.

[11] ISO/IEC 42001:2023, "Information technology - Artificial intelligence - Management system," International Organization for Standardization, 2023.

[12] European Parliament and Council, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," Official Journal of the European Union, 2024.

[13] ETSI GS ZSM 002, "Zero-touch Network and Service Management (ZSM); Reference Architecture," European Telecommunications Standards Institute, 2019.

[14] IEC 61508, "Functional Safety of Electrical/Electronic/Programmable Electronic Safety-related Systems," International Electrotechnical Commission, 2010.

[15] ISO/IEC 27701:2019, "Security techniques - Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management," International Organization for Standardization, 2019.

[16] «XGVela framework, 5G cloud-native open source,» [En línea]. Available: https://xgvela.org/.

[17] M. Pau, M. Mirz, J. Dinkelbach, P. Mckeever, F. Ponci y A. Monti, «A Service Oriented Architecture for the Digitalization and Automation of Distribution Grids, 10.1109/ACCESS.2022.3164393.,» vol. 10, pp. 37050-37063.

[18] «Platform, Open Network Automation,» [En línea]. Available: https://www.onap.org/architecture.

[19] S. Robitzsch, U. Olvera-Hernandez, J. Costa-Requena y M. Skarp, «Enabling Service-Oriented Principles for the User Plane of Mobile Telecommunication Networks,» IEEE Conference on Standards for Communications and Networking (CSCN), 10.1109/CSCN57023.2022.10051037, pp. 111-117, 2022.

[20] K. Katsalis, N. Nikaein, E. Schiller, R. Favraud y T. I. Braun, «5G Architectural Design Patterns,» IEEE International Conference on Communications Workshops (ICC), 2016.

[21] B. Orlandi, S. Lataste, S. Kerboeuf y M. Bouillon, «Intent-Based Network Management with User-Friendly Interfaces and Natural Language Processing,» de 2024 27th Conference on Innovation in Clouds, Internet and Networks (ICIN), France, 2024.
[22] A. Abdellah y A. Koucheryavy, «Survey on Artificial Intelligence Techniques in 5G Networks,» Telecom IT 8.1, pp. 1-10, 2020.
[23] N. Gkatzios, H. Koumaras, D. Fragkos y V. Koumaras, «A Proof of Concept Implementation of an AI-assisted User-Centric 6G Network,» Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), doi: 10.1109/EuCNC/6GSummit60053.2024.10597020, pp. 907-912, 2024.
[24] B. Orlandi, «Intent-Based Network Management with User-Friendly Interfaces and Natural Language Processing,» 27th Conference on Innovation in Clouds, Internet and Networks (ICIN), doi: 10.1109/ICIN60470.2024.10494458, pp. 163-170, 2024.
[25] M. P. Ahokangas, A. Matinmikko-Blue, Basaure y S. Yrjölä, «Use Cases for Local 6G Networks,» oint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), doi: 10.1109/EuCNC/6GSummit60053.2024.10597019, pp. 1127-1132, 2024.
[26] «CONFIDENTIAL6G,» [En línea]. Available: https://confidential6g.eu/.
[27] «DESIRE6G,» [En línea]. Available: https://desire6g.eu/.
[28] «HORSE-6G,» [En línea]. Available: https://horse-6g.eu/.
[29] «PRIVATEER,» [En línea]. Available: https://www.privateer-project.eu/.
[30] «RIGOROUS,» [En línea]. Available: https://rigourous.eu/.
[31] «ELASTIC,» [En línea]. Available: https://elasticproject.eu/.
[32] «iTrust6G,» [En línea]. Available: https://www.sns-itrust6g.com/.
[33] «NATWORK,» [En línea]. Available: https://natwork-project.eu/.
[34] «ROBUST-6G,» [En línea]. Available: https://robust-6g.eu/.
[35] S. Chen, L. Chen, B. Hu, S. Sun, Y. Wang, H. Wang and W. Gao, "User-Centric Access Network (UCAN) for 6G: Motivation, Concept, Challenges and Key Technologies," IEEE Network, vol. 38, no. 3, pp. 154-162, 2024.
[36] e. a. L. Kastner, «On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness,» de IEEE 29th International Requirements Engineering Conference Workshops (REW), Notre Dame, IN, USA, 2021.
[37] L. Chazette, W. Brunotte y T. Speith, «Exploring explainability: A definition, a model, and a knowledge catalogue,» de IEEE 29th International Requirements Engineering Conference (RE), 2021.
[38] W. Pieters, «Explanation and trust: What to tell the user in security and AI?,» Ethics and Information Technology, vol. 13, nº 1, p. 53–64, 2011.
[39] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti y D. Pedreschi, «A survey of methods for explaining black box models,» ACM Computing Surveys, vol. 51, nº 5, p. 1–42, 2019.
[40] M. T. Ribeiro, S. Singh y C. Guestrin, «Why Should I Trust You?': Explaining the predictions of any classifier,» de Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2016.
[41] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kastner, E. Schmidt, A. Sesing y K. Baum, «What do we want from explainable artificial intelligence (XAI)? – A stakeholder perspective on XAI

and a conceptual model guiding interdisciplinary XAI research,» Articifial Intelligence, vol. 296, 2021.

[42] «Ethics Guidelines for Trustworthy Artificial Intelligence, available online at https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai».

[43] «aerOs, Deliverable 2.7,» [En línea]. Available: https://aeros-project.eu/dissemination/deliverables/.

ANNEX 1: GLOSSARY OF TERMS

General Terms

- **Level of trustworthiness (LoTw):** provided to a user is the result of combined actions taken by the cognitive AI coordinator of the 6G system simultaneously at the application plane, core network plane and resource/cloud continuum plane, using the openness and programmability capabilities of these planes. By deploying Trust Functions (i.e. specialized AFs) that interface with these planes, the cognitive coordinator achieves to apply security, privacy, safety, resilience and reliability actions/measures that aim to accommodate the user's intent and realize a user-centric trustworthy service provision over 6G.

- **Trust and Trustworthiness:** Trust is an attitude that a tenant has towards a 6G system. In contrast, trustworthiness is a system property that creates trust to the 6G tenant/user. A user/tenant trusts (or requires a specific level of trust from) a 6G system, because the 6G system is trustworthy. In other words, the trustworthiness of a 6G system contributes to building the trust level of the tenant/user of the specific system. Thus, the more trustworthy the 6G system is, the higher the trust level of the tenant/user will be [36] [36] .

Computing continuum domain – aerOS terminology

- **Infrastructure Element (IE)**: The fundamental building block within aerOS meta operating system. A physical or virtual computing resource providing the necessary processing power, storage capacity, and network connectivity to support containerized workloads and services. Exposes aerOS runtime on top of provided capabilities being thus the minimum execution unit within the IoT-Edge-Cloud continuum.

- **Domain**: A set of one or more IEs, functionally connected and sharing a common instance of aerOS basic services among them, constituting an administrative domain able to be managed and orchestrated by aerOS Meta-OS and thus be part of the IoT-Edge-Cloud continuum.

- **aerOS Runtime**: The operational environment, running on top of each IE, where the containerized workloads, services, and applications are executed and managed. Builds on top of container runtime, enhanced with standard capabilities needed per IE, to be integrated as part of aerOS domain, regarding its state exposure and self-monitoring.

- **aerOS Basic services**: Services running within each aerOS domain, relying on aerOS runtime on top of IEs, realizing functionalities (such as authentication and authorization, among others) that enable integration of domain in the aerOS ecosystem, as part of the IoT-Edge-Cloud continuum, providing smart decisions and orchestration of aerOS workloads execution, targeting most efficient placement within the domain or to other domains across the federated aerOS network.

- **Data Fabric**: Architectural component that automates the integration of data from heterogenous sources and exposes them through a standard interface. Enables the

transparent exchange of information across the continuum using common interpretable information models and thus making data accessible any time, from anywhere. It serves the needs of a "consumer" to either support an internal Meta-OS procedure or an IoT vertical application execution.

- **High-level Orchestrator (HLO)**: First step of the Federated Orchestration in aerOS is realized by the HLO. This element analyses the state of the continuum leveraging the Data Fabric and takes an allocation decision (service deployment spot). This action takes an Intention Blueprint as an input and delivers a Decision Blueprint as output to the HLO.

- **Low-level Orchestrator (LLO)**: Second step of the Federated Orchestration in aerOS is realized by the LLO. This element interprets the Decision Blueprint coming from the HLO and oversees the actual deployment of workloads in the selected IE(s). Being aware of the underlying container management frameworks, it can convert the allocation order into proper deployment. Several LLOs may live in the same domain.

Cloud native and virtualized networking terminology

- **Cloud Native**: Cloud native practices empower organizations to develop, build, and deploy workloads in computing environments (public, private, hybrid cloud) to meet their organizational needs at scale in a programmatic and repeatable manner. It is characterized by loosely coupled systems that interoperate in a manner that is secure, resilient, manageable, sustainable, and observable. Cloud native technologies and architectures typically consist of some combination of containers, service meshes, multi-tenancy, microservices, immutable infrastructure, serverless and declarative APIs. Definition by CNCF.

- **Cloud-native Network Function (CNF)**: CNF is a software-implementation of a function, or application, traditionally performed on a physical device, but which runs inside Linux containers (typically orchestrated by Kubernetes). Inherit all cloud native architectural and operational principles including K8s lifecycle management, agility, resilience, and observability.

- **Service mesh**: A service mesh is a software layer that handles all communication between services in applications. As applications scale and the number of microservices increases, it becomes challenging to monitor the performance of the services. To manage connections between services, a service mesh provides new features like monitoring, logging, tracing, and traffic control. It is independent of each service's code, which allows it to work across network boundaries and with multiple service management systems.

- **Service Communication Proxy (SCP)**: Some or all the following functionalities may be supported in a single instance of an SCP: (i) Indirect Communication, (ii) Delegated Discovery, (iii) Message forwarding and routing to destination NF/NF service, (iv) Message forwarding and routing to a next hop SCP, (v) Communication security (e.g., authorization of the NF Service Consumer to access the NF Service Producer API), load balancing, monitoring, overload control, etc. (vi) Optionally interact with UDR, to resolve the UDM Group ID/UDR Group ID/AUSF Group ID/PCF Group ID/CHF Group ID/HSS Group ID based on UE identity, e.g.

SUPI or IMPI/IMPU. SCPs can be deployed at PLMN level, shared-slice level and slice-specific level. It is left to operator deployment to ensure that SCPs can communicate with relevant NRFs. To enable SCPs to route messages through several SCPs (i.e., next SCP hop discovery), an SCP may register its profile in the NRF. Alternatively, local configuration may be used. Definition by ETSI.

- **Container Network Interface (CNI)**: CNI consists of a specification and libraries for writing plugins to configure network interfaces in Linux containers, along with several supported plugins. CNI concerns itself only with network connectivity of containers and removing allocated resources when the container is deleted. Because of this focus, CNI has a wide range of support, and the specification is simple to implement.

- **Service Function Chaining (SFC)**: A service function chain defines an ordered set of abstract service functions and ordering constraints that must be applied to packets and/or frames and/or flows selected because of classification. An example of an abstract service function is "a firewall". The implied order may not be a linear progression as the architecture allows for SFCs that copy to more than one branch and also allows for cases where there is flexibility in the order in which service functions need to be applied. The term "service chain" is often used as shorthand for service function chain. Definition by IETF.

6G Core Terminology

- **Service Based Architecture (SBA):** The SBA consists of new architecture where the previous 4G network functions such as Mobility Management Entity (MME), Serving Gateway (SGW), Packet Gateway (PGW) and Home Subscriber Server (HSS), Policy and Charging Rules Function (PCRF) are decomposed into more specific network functions with additional web interfaces for interconnection between them.

- **Mobile Packet Core:** Consists of all the network functions required for providing mobile network connectivity and includes but not limited to; Access Mobility Function (AMF), Session Management Function (SMF), User Plane Function (UPF), Unified Data Management (UDM), Authentication Service Function (AUSF), Unified Data Repository (UDR), Network Repository Function (NRF), Network Slice Selection Function (NSSF), Network Slice Access Control Function (NSCAF), Network Data Analytics Function (NWDAF), Network Exposure Function (NEF), Policy Control Function (PCF) Security Edge Protection Proxy (SEPP).

Openness and Programmability terminology

- **Openness:** refers to the creation of a network environment that supports interoperability, transparency, and flexibility. Open architectures and standards are designed to reduce reliance on proprietary hardware and software by promoting open interfaces, modular components, and multi-vendor ecosystems.

- **Programmability**: refers to the ability to dynamically configure, control, and customize network functions and services through software interfaces. Programmability enables 6G

networks to support diverse use cases, such as IoT, autonomous vehicles, and real-time applications, by adjusting network behaviour in real-time to meet unique demands.

- **Monitoring:** refers to the continuous observation and analysis of 6G network performance and user behavior, ensuring the targeted quality offered to the application users. Continuous monitoring systems can monitor the different planes and devices of the 6G ecosystem and provide metrics through frameworks, like the Common API Framework (CAPIF), to ML methods for providing estimations to achieve a specific LoTw.
- **Application Programming Interface (API):** is a set of rules or protocols that enables software applications to communicate with each other to exchange data, features and functionality.

Use-case specific terminology

- **Virtual Reality (VR):** An immersive technology fully immersing users into a digital environment. This is achieved by making users wear an opaque headset to cut them from their real environment.
- **Augmented Reality (AR):** An immersive technology which consists in adding virtual elements into the real environment of the user. This can be achieved in three major ways: through a handheld device (a tablet for instance), through a head-worn device (for instance a headset) or with an external projection system.
- **Extended Reality (XR):** XR refers to the spectrum of immersive technologies that mix virtual and physical elements. AR and VR are among the most well-known representatives of such technologies but do not represent the whole spectrum. XR involves a large panel of potential devices, from smartphones and headsets for AR to complete room-scale systems like CAVEs. XR can be used as a synonym of Mixed Reality (MR).
- **Digital Twin (DT):** A DT refers to the digital counterpart of a given system or process. This digital element reflects with a high degree of fidelity its physical counterpart, allowing simulation, monitoring and maintenance operations on it.
- **Metaverse**: Metaverses are persistent and shared set of interaction spaces that exist both locally and remotely. Metaverses involve a lot of different technology to enable these persistent virtual spaces, including but not limited to XR, blockchains, digital twins and IA. Many definitions, viewpoints and confusion still co-exist about metaverses since it is both a recent and broad concept.

MLOps / DataOps terminology

- **Artificial Intelligence (AI)**: Discipline to create machines with the ability to perform some cognitive functions usually associated with human minds, such as perceiving, reasoning, learning, interacting with the environment, problem solving, and even exercising creativity.

- **Machine Learning (ML):** Subfield of AI that, using data and algorithms, provides machines with the capability of learning the way humans do, without specific programming, gradually improving its accuracy.
- **Deep Learning (DL):** Type of ML that uses artificial neural networks to learn from data, as opposed to classical ML based on Decision Trees, Support Vector Machine (SVM), etc.
- **Federated/Collaborative Learning:** Type of ML that uses collaborative data provided by multiple entities but ensuring that their data remains decentralized and there is no exchange of data from client devices to global servers as the learning process takes place locally, increasing data privacy.
- **ML models:** Computer programs or systems created from ML algorithms used to recognize patterns in data or make predictions.
- **Trained ML models:** ML models that once trained can make useful predictions from new input data.
- **Supervised ML models:** ML models that use labeled data as input.
- **Unsupervised ML models:** ML models that use unlabeled data as input.
- **Reinforcement ML models:** ML models that learn by receiving feedback about performance after deployment.
- **Predictive or discriminative ML models:** ML models used for classifying or predicting data labels.
- **Generative ML models:** ML models able to generate new data / content from input data.
- **ML Operations (MLOps):** Process of managing the ML life cycle, from development to deployment.
- **Dataset**: In the context of Artificial Intelligence, a dataset is used to train learning techniques to reproduce a given behavior. We distinguish input data (provided by the system) from expected data (that we want to predict). To be used for learning, a dataset must pair inputs with expected outputs.
- **Synthetic data:** Data artificially created by computer algorithms, as opposed to real data that are collected from real events.

Chatbot terminology

- **Natural language processing (NLP):** NLP refers to the computational capability to comprehend and interpret human language. It involves recognizing patterns in communication and converting written text into spoken language.

- **Natural language understanding (NLU):** NLU is the ability of a computer system to comprehend the meanings and sentiments in natural language, facilitating a deeper understanding of human expressions.
- **Natural Language Generation (NLG)**: NLG allows the chatbot to generate human-like text responses, improving the fluidity of communication and making the interaction feel more natural and engaging.
- **Conversational AI (CAI)**: A form of artificial intelligence that powers the ability of chatbots to engage in natural language conversations with users. This allows chatbots to simulate human-like interactions, improving user experience by understanding and responding to queries in real time.
- **Sentiment analysis:** Sentiment analysis enables the evaluation of the tone and sentiment in written or spoken language. Through intent detection, chatbots and similar systems can analyze messages to assess the user's sentiment—whether positive, negative, or neutral—toward a product or service.
- **Intent detection**: Intent detection refers to the process of identifying a user's purpose or intention based on their communication.
- **Intent Classification**: Part of intent recognition, this is where the chatbot categorizes the user's input into specific predefined intents to trigger the appropriate action or response.
- **Response Generation**: The process where the chatbot formulates a response based on user input. This can involve retrieving information from databases, using AI to craft a response, or running processes like intent recognition to give meaningful answers.
- **Fallback Mechanism**: A strategy used when the chatbot fails to understand the user's intent. It provides predefined generic responses or escalates the conversation to a human agent to ensure the conversation continues smoothly.

Cognitive Coordinator terminology

- **Regression Component:** A component that uses AI algorithms, like logistic regression, to convert textual data into numerical values for analysis and processing.
- **Queries:** Requests for specific information from a system or database.
- **Knowledge Base:** A structured database, such as MongoDB or SQL, that stores organized user and system information for easy retrieval and analysis.
- **Reasoning Engine:** A component that retrieves relevant information from a knowledge base by executing targeted queries and utilizes them to refine trustworthiness scores.

XAI terminology

- **Tabular data:** In this project, we will focus on so-called tabular data for inputs & outputs of AI techniques. Tabular data represents numerical or categorical values that are organized in a table. We oppose tabular data to signal data (such as images, audio files, video files…) which are unstructured. In tabular data, each input/output is structured as column for which we have a clear interpretation of the semantics.

- **Model (AI context):** In the context of Artificial Intelligence, a model is a mathematical object that can be parameterized to fit a given relationship between input data and expected outputs. Finding the right parameters for the model is computing intensive and rely on a dataset of examples that we aim to generalize. The term model encompasses both the method (various mathematical models exist, including but not limited to neural networks) and the hyperparameters learned to best fit the example dataset.

- **Explainable AI (XAI)**: XAI is a domain of Artificial Intelligence dedicated to providing explanations to end users and / or ML engineers as to the rationales behind a decision taken by a ML algorithm. This encompasses both statistical and neural net approaches.

- **Large Language Models (LLMs):** LLMs are a category of Neural Nets relying on the Transformers architecture applied to natural language understanding tasks. It is now a popular set of models to tackle human machine interaction & cooperation, including in decision taking contexts.

- **Local Explainability:** In Explainable AI, a local explanation focuses on understanding the decision of a model for a specific instance. It explains why the model made a particular prediction for an individual input.

- **Global explainability:** In Explainable AI, global explanations provide an overview of how the AI model generally works. It describes the overall patterns and factors that influence the model's decisions across all instances.

Security Trust Function Terminology

- **Self-Sovereign Identity (SSI):** A digital identity model where individuals have complete ownership and control over their personal data. In an SSI framework, users can store their identity information in secure digital wallets and share it selectively with service providers. This approach minimizes reliance on centralized authorities, enhancing privacy and security by preventing unauthorized access and reducing the risk of data breaches.

- **Decentralized Identifier (DID):** A globally unique identifier that enables verifiable, decentralized digital identities. DIDs are not tied to any centralized registry or authority and are often recorded on distributed ledger technologies like blockchain. They allow individuals and entities to establish secure, private connections and control their identifiers independently, facilitating self-sovereign identity management.

- **Verifiable Credentials (VC):** A tamper-evident, digital credential that can be cryptographically verified. Verifiable credentials are issued by trusted authorities and contain information about an individual or entity. They enable secure sharing of credentials (e.g., passports, diplomas, certifications, etc.) in a way that the recipient can trust the authenticity and integrity of the information without exposing unnecessary personal data.

- **Intrusion Detection System (IDS):** A security solution designed to monitor network or system activities for malicious actions or policy violations. An IDS analyzes traffic patterns to detect suspicious activities, such as unauthorized access attempts or malware. Upon detection, it alerts administrators but does not take direct action to block the threats, serving as an early warning system for potential security incidents.

- **Intrusion Prevention System (IPS):** An advanced security mechanism that not only detects potential threats like an IDS but also takes proactive measures to prevent them. An IPS sits in line with network traffic and can automatically block or reject malicious activities in real-time by dropping packets, terminating connections, or configuring firewalls, thereby actively protecting the network from attacks.

- **Smart Contracts:** Self-executing contracts with the terms of the agreement directly written into code. Smart contracts automatically enforce and execute the agreed-upon rules and obligations when predefined conditions are met, without the need for intermediaries. They operate on blockchain platforms, ensuring transparency, immutability and security in transactions ranging from financial exchanges to supply chain management.

- **Blockchain:** A decentralized, distributed ledger technology that records transactions across a network of computers. Each block contains a list of transactions and is linked to the previous one using cryptography, forming a chain. Blockchain ensures data integrity, transparency and security by making it virtually impossible to alter past records without consensus from the network. It's the foundational technology behind cryptocurrencies like Bitcoin and enables various applications like smart contracts and decentralized apps.

Safety Trust Function Terminology

- **Software Defined Perimeter GateWay (SDPGW):** SDPGW is the dynamically created function Gateway to redirect user traffic to specific/target applications.

- **Software Defined Perimeter Control Function (SDPCF):** SDPCF is the dynamically created function that is going to manage gateway and user connectivity lifecycle.

- **AiA:** Nested AI Agent is responsible for decision making and communicating with cognitive coordinator. The specific policies that are going to be applied to the UE's connectivity derive from the output of the AI agent.

- **Authentication and Authorization (AnA):** AnA is a subfunction of Safety AF, which is responsible for the proper authentication of UE/device and communicating with Network

Functions such as NWDAF, UDM and PCF to collect the user information and create applicable model for requested service.

- **Function Lifecycle Management (FLCM):** FLCM is a subfunction of Safety AF, which is responsible to communicate with MANO/aerOS to request the action to deploy or remove the dynamically created SDPGW & SDPCF.

Privacy Trust Function Terminology

- **Privacy:** Privacy refers to the protection of personal data and communication information from unauthorized access or misuse in 5G networks and services.
- **Level of Privacy (LoP):** The **LoP** that can be guaranteed to the specific service at a given time.
- **Desired Level of Privacy (dLoP):** The **LoP** that the **Cognitive Coordinator** requested for a specific service.
- **Final Level of Privacy (fLoP):** The maximum **LoP** that can be achieved by the network at a given time for a given service. It is only calculated and reported to the **Cognitive Coordinator** when the **dLoP** cannot be satisfied.
- **Privacy Action:** An action that can be evoked by the **Privacy Function** and affects the **LoP** of the given service.
- **Decision Support System (DSS):** The **DSS** is responsible for deciding the **Privacy Actions** that need to be taken in order to achieve a higher **LoP** for the service.

Resilience Trust Function Terminology

- **Resilience**: refers to the capability of the network to sustain service quality and reliability through efficient resource management, customizing them from user-intents and adapting to variations in network loads and infrastructure conditions to ensure continuous service availability.
- **Level of Resilience (LoR)**: refers to the resilience level that can be provided to a specific service at a given moment, based on current network resource allocation and infrastructure capacity.
- **Desired Level of Resilience (dLoR)**: refers to the resilience level requested by the Cognitive Coordinator based on an input from LoTw for a service to handle expected changes in network resource demand and infrastructure load.
- **Resilience Action:** refers to an action triggered by the Resilience Function to apply, adjust or improve the LoR of a user for a specific service, such as prioritizing, rerouting traffic or allocating additional resources.

Reliability Trust Function Terminology

- **Reliability:** refers to the ability of the 6G network to consistently perform its intended function under predefined conditions, ensuring the targeted quality offered to the application users. Various reliability-related mechanisms are offered by the SAFE-6G to provide a high LoTw, including data collection from different planes and devices, training and deployment of ML methods, as well as generation of alerts.
- **Alerts:** refer to mechanisms designed to disseminate critical information quickly and efficiently during cases of significant events. Examples of significant events include Quality of Experience (QoE) degradation, reaching system breaking points, detection of abnormal operations, etc. The alerts leverage the capabilities of SAFE-6G to ensure the LoTw.
- **Service Profiling** refers to the process of defining, analyzing, and optimizing the various services supported by the 6G network. This involves analyzing and understanding the specific requirements of different applications and user scenarios to ensure that the network can deliver the necessary performance, reliability, quality of service (QoS) and QoE to the application users. The service profiling aids in identifying patterns and potential bottlenecks that could negatively impact reliability. Service profiling also aids in assessing how new technologies might introduce vulnerabilities to the 6G system, allowing for taking preemptive measures.