6G AI-Native Integrated RAN-Core Networks

Deliverable D2.3
6GARROW Refined Scenarios, Use Cases and related KPIs/KVIs

Co-funded by the European Union

This work has been supported by the 6GARROW project which has received funding from the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101192194 and from the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00435652).

Date of delivery:	29/05/2026	Version:	1.0
Project reference:	101192194	Call:	HORIZON-JU-SNS-2024-STREAM-B-01-06
Start date of project:	01/01/2025	Duration:	36 months

Document properties:

Document Number:	D2.3
Document Title:	6GARROW Refined Scenarios, Use Cases and related KPIs/KVIs.
Editor(s):	Zoran Utkovski (FhG), Davide Montagno Bozzone (HPE)
Authors:	Arman Ahmadian (Yonsei), Chengjie Ma (Yonsei), Yeosun Kyung (Yonsei), Nicolas Cassiau (CEA), Hannu Flinck (Aalto), Zoran Utkovski (FhG), Yuzhen Ke (FhG), Louis-Adrien Dufrène (Orange), Quentin Lampin (Orange), Taeyeon Kim (ETRI), Hoon Lee (ETRI), Davide Montagno Bozzone (HPE), Valerio Frascolla (INT)
Contractual Date of Delivery:	29/05/2026
Dissemination level:	R-PU (Public)
Status:	Draft
Version:	1.0
File Name:	6GARROW D2.3

Abstract

This document refines the use cases introduced in D2.1 with emphasis on newly emerged aspects and novelties with respect to the State of the Art. Building on the scenarios introduced in D2.1 and the functional architecture established in D2.2, D2.3 identifies functional and non-functional requirements, technical enablers, KPIs, KVIs, and potential ethical risks for each of the studied use cases. The document further provides a gap analysis against current 3GPP systems. Finally, it provides a mapping of the use cases to the initial 6GARROW functional architecture introduced in D2.2 and discusses relations to the demonstration activities of the project.

Keywords

6G, Use Cases, Use Case Families, Key Performance Indicators (KPIs), Key Value Indicators (KVIs)

Disclaimer

Executive Summary

Deliverable D2.3 advances the 6GARROW project's work on Artificial Intelligence (AI)-native network architecture by refining the use cases introduced in D2.1 and aligning them with the functional architecture developed in D2.2. D2.3 provides a high-level link between application-driven needs, architectural development, AI-related performance assessment, and future validation activities.

The deliverable focuses on the following key aspects:

- **AI-related KPIs and KVIs:** Key Performance Indicators (KPIs) and Key Value Indicators (KVIs) for 6G are introduced to quantify operational efficiency and commercial value from both semantics-oriented AI and syntax-oriented AI perspectives.
- **Use case Refinement:** Selected use cases from D2.1 are revisited with updated requirements, gaps with respect to current 3GPP systems, technical enablers, KPIs, KVIs, ethical risks and new perspectives, where applicable.
- **Architectural Mapping:** The refined use cases are mapped to the functional architecture defined in D2.2 and linked to the project's demonstration activities, supporting alignment between use case analysis, architectural design, and future Proof-of-Concept (PoC) validation.

List of Tables

List of Figures

Acronyms and abbreviations

Term	Description
5GC	Fifth-Generation Core
AE	Autoencoders
AI	Artificial Intelligence
AI/ML	Artificial Intelligence / Machine Learning
AIaaS	Artificial Intelligence as a Service
AMF	Access and mobility Management Function
API	Application Programming Interface
ARIMA	Auto-Regressive Integrated Moving Average
A2A	Agent-to-Agent
Bi-LSTM	Bidirectional Long-Short Term Memory
B5G	Beyond 5G
CAPEX	Capital Expenditure
CMC	Core Management Core
CN	Core network
CNN	Convolution Neural Network
CPU	Central Processing Unit
CSI	Channel State Information
C-plane	Control-plane
DDPG	deep deterministic policy gradient
DL	Deep Learning
DoS	Denial of Service
DRA	Dynamic Resource Allocation
DRL	Deep Reinforcement Learning
DT	Digital Twin
EE	Energy Efficiency
EEA	European Economic Area
E2E	End-to-End
ETSI	European Telecommunications Standards Institute
EU	European Union
GAE	Graph Autoencoder
GDRP	General Data Protection Regulation
GenAI	Generative AI
gNB	Next generation Node B
GNN	Graph Neural Network

GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
GUI	Graphical User Interface
HLM	Hybrid Language Model
IBN	Intent Based Networking
ID	Identifier
IMEI	International Mobile Equipment Identity
IoT	Internet of Things
ISAC	Integrated Sensing And Communication
JCAS	Joint Communications And Sensing
KPI	Key Performance Indicator
KVI	Key Value Indicator
LLM	Large Language Model
LMF	Location Management Function
LPWAN	Low-Power Wide-Area Network
LSTM	Long Short-Term Memory
M2M	Machine-to-Machine
MAE	Mean Absolute Error
MCP	Model Context Protocol
MDA	Management Data Analytics
MEC	Multi-Access Resource Allocation
MIMO	Multiple Input Multiple Output
ML	Machine Learning
MNO	Mobile Network Operator
MRSS	Multi-Radio Spectrum Sharing
NAS	Neural Architecture Search
NB-IOT	Narrow Band Internet of Things
NEF	Network Exposure Function
NF	Network Function
nGRG	Next Generation Research Group
NLP	Natural Language Process
NR-U	New Radio Unlicensed
NW	Network
NWDAF	Network Data Analytics Function
OAM	Operations, Administration and Maintenance
OFDMA	Orthogonal Frequency-Division Multiple Access
OPEX	Operational Expenditure
O-RAN	Open RAN

PDU	Protocol Data Unit
PHY	Physical Layer
PoC	Proof-of-Concept
PPO	Proximal Policy Optimization
PRB	Physical Resource Blocks
QoE	Quality of Experience
QoS	Quality of Service
RA	Resource Allocation
RAN	Radio Access Network
RAT	Radio Access Technology
RGB-D	Red-Green-Blue-Depth
RIC	RAN Intelligent Controller
RL	Reinforcement Learning
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
ROI	Return on Investment
RS	Reference Signal
RT	Real Time
SAC	Soft Actor-Critic
Seq2Seq	Sequence to Sequence
SI	Study Item
SLG	Service-Level Guarantee
SLM	Small Language Model
SMF	Session Management Function
SSLA	Semantic Service Level Agreement
SSRF	Semantic State Representation Function
S-RIC	Semantic Radio Controller
TCO	Total Cost of Ownership
TD3	Twin-Delayed Deep Deterministic
TFT	Temporal Fusion Transformer
TR	Technical Report
UAV	Unmanned Aerial Vehicle
UE	User Equipment
UPF	User Plane Function
URLLC	Ultra-Reliable Low Latency Communications
U-HLM	Uncertainty-aware opportunistic Hybrid Language Model
VAE	Variational autoencoder
VPN	Virtual Private Network
VLA	Vision Language Action
Wi-Fi	Wireless Fidelity

WWAN	Wireless Wide-Area Network
XR	Extended Reality

Table of Contents

1 Introduction

The present document details refinement of the key 6GARROW use cases introduced in D2.1 “6GARROW Scenarios, Use Cases and related KPIs/KVIs” [6GARROW-D21], clustered according to use case Families as illustrated in Table 1 (from D2.1).

The deliverable focuses on the following key aspects:

(i) revisit AI-related KPIs and KVIs, with the objective to quantify operational efficiency and identify potential commercial value of the studied use cases;
(ii) selected use cases from D2.1 are revisited with updated requirements, gaps with respect to current 3GPP systems, technical enablers, KPIs, KVIs, ethical risks and new perspectives where applicable;
(iii) map the refined use cases to the functional architecture defined in D2.2, connecting to the project’s demonstration activities.

Table 1: 6GARROW Use Case Families and Use Cases (from D2.1).

Families	1. AI native 6G design empowering new network services for stakeholders	AI native 6G design for substantially more efficient & sustainable network operation:	
		2. Resource Usage (lower layers)	3. Automation & Failure Recovery (network functions)
Use Cases	AI-as-a-Service for stakeholders & creation of tailored AI models according to End Users needs	Highly efficient communications with light infrastructure	Anomaly Detection & Recovery Strategies
	URLLC for Machine Vision-based Robotic Control	Dynamic resource allocation for intercontinental RAN-Core	Cultivating Radio Measurements into Value-Added Services for Spectrum Sharing
	Semantic Communications Services for Agents	Overhead reduction through Semantic Closed Control Loop for Private 5G Management Systems	
		RAN/UE & RAN/CN Cross-Domain Coordination	

We note that, in the above terminology, the term “AI native 6G” network design is used in the sense that the network is defined and optimized for AI processing, relying on low-latency and ubiquitous access to intelligence. This is typically not the case for 5th (5G) and earlier generations of mobile networks, where AI functionalities are added “on-top” of a well-advanced network design.

Another useful way to understand the 6GARROW use cases is to look at them from two perspectives: what AI is used for, and what kind of communication problem the use case addresses. Some use cases treat AI as a service offered by the network to users, applications, or external stakeholders. Others use AI internally to improve how the network operates, for example by optimizing resources, reducing overhead, or recovering from failures. At the same time, some use cases are ***semantics-oriented***, meaning they focus on meaning, intent, context, or agent interaction, while others are

syntax-oriented, meaning they focus on conventional network metrics such as latency, throughput, reliability, and resource allocation. The chart below categorizes the use cases according to these two dimensions while using colors to denote the use case family they belong to.

Table 2: Classification of 6GARROW use cases by AI role, communication orientation, and objective.

<table>
<tr><th></th><th>Semantics-oriented</th><th>Syntax-oriented</th></tr>
<tr><td rowspan="4">AI-on-Network</td><td>Semantic communications services for agents</td><td rowspan="3"></td></tr>
<tr><td>URLLC for machine vision-based Robotic control</td></tr>
<tr><td>Highly efficient communications with light infrastructure</td></tr>
<tr><td colspan="2">AI-as-a-Service for stakeholders & creation of tailored AI models according to end-user needs</td></tr>
<tr><td rowspan="4">AI-for-Network</td><td rowspan="2">Overhead reduction through semantic closed control loop for private 5G management system</td><td>Dynamic resource allocation for intercontinental RAN-Core</td></tr>
<tr><td>Anomaly detection and recovery strategies</td></tr>
<tr><td colspan="2">RAN/UE & RAN/CN cross-domain coordination</td></tr>
<tr><td colspan="2">Cultivating radio measurements into value-added services for spectrum sharing</td></tr>
</table>

2 Methodology & Structure

Combined with D2.1 “6GARROW Scenarios, Use Cases and related KPIs/KVIs” [6GARROW-D21] and D2.2 “Initial Architecture of 6GARROW AI-native Network” [6GARROW-D22], this deliverable establishes a coherent, iterative workflow that builds upon high value 6G application scenarios to drive architectural evolution, requirement setting, and technology validation. The applied use case-centric approach begins with anticipated real-world demands, focusing on domains like mission-critical communication, achievable through semantic information processing, and intelligent automation. The identified use cases are grouped into three families based on functional and business alignment. Each family has been assessed for its potential to drive architectural shifts within the 6GARROW framework. The use case families are used for systematically derive architectural requirements, performance metrics, and AI/ Machine Learning (ML) integration strategies from high-value application scenarios. This deliverable refines the scenarios and use cases provided in D2.1 with the emphasis on the new aspects not yet covered in that deliverable.

For each use case, functional and non-functional requirements with technical enablers, such as MLOps pipelines, access to essential data, and meaning extraction for semantic encoding, are identified. The focus is on how AI and semantic processing can be introduced, for example to reshape the network control loops, reduce signaling overhead, or enhance responsiveness. Similarly, for each use case, the associated KPIs and KVIs are documented. Furthermore, possible ethical risks are identified and documented.

The methodology includes a forward-looking gap analysis that compares the state of the art in current 3GPP systems with the disruptive capabilities of AI-native, semantic-enabled, and agent-based architectures envisioned for 6G.

Derivation of AI relevant KPIs and KVIs, that are discussed in detail in the subsequent sections, are derived to quantify both operational efficiency (e.g., reduced latency, energy savings) and commercial value (e.g., return on investment, infrastructure cost reduction) resulting from the use of native AI. In chapter 4, mapping of the use cases against the functional architecture developed in D2.2 is presented as well as how the use cases relate to the prototyping efforts of the project.

2.1 Feedback of the Advisory Board

The project’s Advisory Board was kindly asked to review the 6GARROW use cases described in D2.1. The obtained feedback was then used to refine the use cases in this deliverable, focusing on two main areas: technical feasibility/credibility and business perspective.

Regarding technical feasibility, the Advisory Board noted that because this project is an early-stage exploration, strict technical feasibility is less critical right now. Furthermore, hardware performance for AI inference is improving rapidly in both capacity and efficiency, which will support these use cases. However, the Advisory Board raised concerns about security, reliability, and privacy regarding AI/semantics at the physical (PHY) layer. Since this is a new topic, such questions are expected. To address this, we have added a dedicated risk section for each use case.

Regarding the business perspective, vertical industries are primarily interested in the return on investment (ROI) of deploying AI-native networks. While it is too early to calculate the exact financial value of AI, we have included a business case rationale for each use case to help verticals identify potential market opportunities.

2.2 KPIs and KVIs

In this section, we introduce the KPIs and KVIs that are expected to characterize 6G communication systems.

KPIs quantify technical performance, including latency, reliability, and efficiency, whereas KVIs capture broader value dimensions, such as sustainability, service utility, trustworthiness, and economic benefit.

We have identified only two KVIs in the scope of 6GARROW: ROI and trust. All other itme discussed in this deliverable are covered by the KPI definition above.

Several classical KPIs remain essential and are summarized in Table 3 together with representative values reported in literature [6GIA24], [NGA24], [GovI23], [NICT22], [B5G22], [SKT23a], [SKT24], [IMT23a], [IMT23b], [TAICS23], and [M.2410-0].

Table 3: Summary of regional and global KPI and KVI targets

KPIs	Unit	Networld Europe SRIA 2022	Next G Alliance (USA)	IMT-2030 PG (China)	B5G Consortium (Japan)	TSDSI (India)	TAICS (Taiwan)	ITU IMT-2030
Peak Data Rate	**Tb/s**	1	0.5-1	1	0.1–0.2	0.5–1	0.1~1	0.50~0.200
User Experienced DL Data Rate	**Gb/s**	10	≤1	10–100	10–100	≤10	1	0.3~0.5
User Experienced UL Data Rate	**Gb/s**	10	≤1	10–100	10–100	≤5	1	0.3~0.5
Density	**UE/m2**	1	1	1	1	1	1	1~100
Reliability [BLER]	**Number of nines**	8	8	7	7	7	5	5 to 7
U-Plane Latency	**ms**	<0.1	0.1 ~ 1	0.1	0.1 ~ 1	0.1 ~ 1	0.1	0.1 ~ 1
Network Energy Efficiency (EE)	—	>100% gain vs IMT-2020	Extremely low power	100x w.r.t 5G	100x w.r.t 5G	Battery lifetime up to 20 years	10% of 5G	N/A
Terminal EE	—	>100% gain vs IMT-2020	Never charging devices	20 years	N/A	Up to 20 years battery lifetime	20 years battery lifetime	N/A
Mobility	**km/h**	<1000	>500	N/A	N/A	<=1000	<= 1000	500~1000
Indoor Positioning Accuracy	**cm**	<1	0.1 ~ 10	1	1 ~ 2	<1	10	1 ~ 10
Outdoor Positioning Accuracy)	**cm**	<1	0.1 ~ 10	50	1 ~ 2	<1	N/A	1 ~ 10

The rest of this section examines how AI influences the KPIs and KVIs identified for 6G, while distinguishing between two fundamentally different modes of impact (as discussed in Chapter 1 and summarized in Table 2).

On the one hand, **semantics-oriented AI** changes the communication paradigm itself by focusing on task-relevant meaning rather than exact bit delivery, which may require some conventional metrics to be reinterpreted or complemented by outcome-based views. On the other hand, **syntax-oriented AI** operates within the classical communication framework and enhances existing performance metrics through improved prediction, optimization, scheduling, and resource management. By separating these two perspectives, this section aims to clarify where AI modifies the meaning of performance assessment, where it directly improves conventional indicators, and how both contribute to the broader AI-native vision of 6G.

We first include below a set of secondary KPIs and KVIs that are either difficult to quantify or have been studied to a lesser extent in the literature (e.g., due to some of these being of secondary technical significance), such that there is no consensus on the values to be reported.

In relation to **syntax-oriented AI**, we consider as relevant the following KPIs and KVIs:

- **Autoregressive integrated moving average (ARIMA) mean absolute error (MAE)**: Measures the average error between predicted and actual values in time-series forecasting of network traffic and resource demand, reflecting accuracy in numerical prediction models;
- **ARIMA root mean square error (RMSE)**: Quantifies prediction error with higher sensitivity to large deviations, evaluating the robustness of traffic and load forecasting models used for network optimization;

- **User experience measurement**: Behind AI models and agent communications lies applications connected to user requests or services. Semantic communications and AI-native 6G networks should configure equipment and services with application-specific metrics as primary considerations. This user-centric approach ensures that technical optimizations directly translate to tangible improvements in service quality, responsiveness, and overall satisfaction, creating a virtuous cycle of adoption and utilization;
- **Return of Investment (ROI)**: Implementation of agent communications and dedicated services such as Compute-as-a-Service (CaaS) or AI-as-a-Service (AIaaS) represents new monetization streams for operators while leveraging a natural evolution from software-based 5G architecture. These services should be introduced incrementally, allowing mobile network operators (MNOs) to amortize investments while gradually expanding capabilities and service offerings that capitalize on existing infrastructure investments;
- **Trust**: A significant concern surrounding AI and agents is the potential for data leakage and privacy violations. As service providers operating on their networks, MNOs bear responsibility for client data protection. This underscores the necessity for comprehensive security measures including data privacy safeguards, robust authentication mechanisms, and granular authorization controls. Additionally, AI agents must be properly authenticated on the network, while the network itself must be capable of verifying legitimate AI agent services to maintain system integrity and user confidence.

However, 6G is not merely a faster version of 5G; rather, it is envisioned as an AI-native system with semantic communication, integrated sensing, and distributed intelligence capabilities. These features challenge purely conventional performance metrics. AI influences KPIs and KVIs in two distinct ways: by enhancing conventional communication processes and by reshaping the communication paradigm itself through semantics-oriented AI. Therefore, a shift in perspective is needed when evaluating 6G performance.

The following represents some KPIs directly related to **semantics-oriented AI**:

- **RMSE semantic state accuracy**: Measures the deviation between raw network data (e.g., traffic, sessions, signaling events) and their semantic representations generated by semantic state representation function (SSRF). It captures how accurately low-level operational information is translated into meaningful semantic descriptions of the network state;
- **Semantic similarity (SSM)**: Evaluates how closely the meaning of original network descriptions matches their AI-generated semantic representations using embedding-based comparison in a high-dimensional semantic space;
- **Answer correctness (AC)**: Assesses the quality of AI-generated interpretations by combining factual correctness with semantic similarity, reflecting alignment with expert-level understanding of network conditions;
- **Natural language processing (NLP) F1-score**: Measures the effectiveness of semantic classification of network alerts, balancing precision and recall in identifying different network states;
- **NLP precision**: Evaluates how accurately the system classifies detected network conditions, minimizing incorrect semantic interpretations;
- **NLP recall**: Measures how completely relevant network conditions are identified from textual or log-based inputs;
- **Compression rate**: Closely tied to latency objectives, the semantic communications field aims to efficiently compress representations exchanged between AI models. Effective compression algorithms must preserve task integrity and model interoperability while reducing transmission overhead. The ideal compression approach balances semantic fidelity with processing requirements, ultimately benefiting both system responsiveness and energy consumption.

2.2.1 Assessing the impact of semantics-oriented AI on KPIs/KVIs

The introduction of semantics-oriented AI changes the way network performance should be interpreted and assessed. Unlike conventional communication, which primarily optimizes the accurate delivery of bits, semantics-oriented communication seeks to deliver the information that is most relevant to the intended task or service outcome. As a result, its effect on KPIs and KVIs is not uniform: some indicators may benefit directly from reduced payload and more selective transmission, while others are only weakly affected or may require reinterpretation. In this section, we therefore examine each KPI and KVI individually, with the goal of identifying where semantics-oriented AI offers genuine performance gains, where its impact is mainly indirect, and where conventional metrics should be complemented by outcome-based views.

For the reasons mentioned above, in the following we provide more details on the KPIs from Table 3:

- **Peak data rates**: Apart from the fact that semantics-oriented AI can potentially reduce peak data-rate requirements for the same task, the impact on peak data rate is negligible;
- **User experienced data rate**: The impact can be significant, but interpretation becomes more subtle. In a strict conventional sense, the user may consume fewer bits because less information is sent. Yet from a service perspective, perceived performance may improve because the same task or experience is achieved with less traffic. As a result, this KPI may need reinterpretation through an equivalent or outcome-based view;
- **Density**: The impact can be positive in systems with many devices or agents. If each device transmits only semantically relevant information, the network can support a larger number of active users within the same resource budget. This is particularly relevant for dense Internet of Things (IoT), sensing, or distributed AI scenarios;
- **Reliability**: Semantics-oriented AI can improve effective reliability when the service objective is task completion rather than exact bit recovery. By transmitting only the most relevant information, the system may remain operational even under imperfect channel conditions. However, if semantic extraction or reconstruction fails, the impact can be severe, so reliability must be evaluated at both the communication level and the task level;
- **User-plane (U-plane) latency**: The impact on U-plane latency is often significant and positive. Since semantic communication reduces the amount of transmitted payload, transmission and processing delays can decrease, especially for vision, sensing, and control applications (e.g. see Section 3.2.2). That said, gains may be partly offset by AI inference time at the sender and receiver;
- **Network Energy Efficiency (EE)**: The impact can be positive if semantic processing substantially reduces traffic volume and repeated retransmissions. However, this benefit must be balanced against the extra computation required for semantic encoding, decoding, and model execution. The resulting energy costs can be partially mitigated through techniques such as quantized ML, which reduce computational complexity and power consumption;
- **Terminal EE**: On the one hand, transmitting less raw data can reduce energy consumption, which is particularly advantageous for uplink-intensive applications. On the other hand, performing local AI inference for semantic extraction introduces additional computational load and may increase battery usage. As a result, the overall impact depends on device capabilities and model complexity. When on-device inference is impractical or incurs excessive hardware or energy costs, split inference approaches, such as those described in Section 3.1.2, become important. In this context, applying model optimization techniques, including quantized ML, is critical for reducing energy consumption at the terminal;
- **Mobility**: The direct impact on mobility is usually small. Semantics-oriented AI does not fundamentally change the maximum supported speed of a user or device. Still, it may improve

mobility robustness indirectly by reducing traffic overhead during handovers or by prioritizing only task-relevant updates in highly dynamic scenarios;

- **Indoor and outdoor positioning accuracy**: The direct impact is usually limited. Semantic-oriented AI does not generally improve the classical positioning-accuracy KPI itself, since this still depends mainly on radio measurements and sensing capability. Its contribution is mostly indirect: by reducing data traffic, semantic communications can free up radio resources that may then be allocated to location-related sensing and data-gathering functions.

2.2.2 Assessing the impact of syntax-oriented AI in enhancing KPIs/KVIs

Syntax-oriented AI enhances network performance by improving the transmission, control, and management of conventional data flows without changing the fundamental bit-oriented nature of communication. Unlike semantics-oriented AI, which focuses on task-relevant meaning, syntax-oriented AI operates within the classical communication framework and aims to improve existing KPIs and KVIs through better prediction, optimization, scheduling, resource allocation, and control. Its impact is therefore often more direct on the (more) traditional network metrics from Table 3:

- **Peak data rates**: The impact on peak data rates is usually limited but not negligible. Syntax-oriented AI can improve beamforming, Modulation and Coding Scheme (MCS) selection, interference mitigation, and channel estimation, which may help approach the theoretical peak more often. However, it does not fundamentally redefine the physical limits set by spectrum, bandwidth, and hardware;
- **User Experienced Data Rate**: The impact on user experienced data rate can be significant. Syntax-oriented AI can improve fairness, scheduling, congestion control, and radio RA so that more users experience stable and higher effective throughput in realistic conditions. This KPI is often more sensitive to intelligent optimization than peak data rate, because it reflects practical network behaviour rather than ideal conditions;
- **Density**: The impact on density can be significant in highly loaded systems. Syntax-oriented AI can support more simultaneous devices by optimizing random access, scheduling, interference coordination, admission control, and traffic prioritization. This is especially relevant in IoT and massive access scenarios where the challenge is efficient coordination of many conventional data flows;
- **Reliability (Number of nines)**: Syntax-oriented AI can have a strong positive impact on reliability by improving conventional communication functions such as link adaptation, channel prediction, scheduling, retransmission control, and anomaly detection. By predicting poor channel conditions and proactively adjusting resources, the network can reduce packet loss and service interruptions. In this case, reliability is improved in the classical sense of successful bit delivery, without changing the meaning of the metric;
- **U-plane latency**: The impact on U-plane latency can be significant. Syntax-oriented AI can reduce delay through predictive scheduling, queue management, traffic steering, congestion avoidance, and faster adaptation of radio resources. Unlike semantics-oriented AI, the gain does not come from reducing payload size, but from making the transport of conventional data streams more efficient;
- **Network EE**: The impact on network EE can be substantial. Syntax-oriented AI can optimize sleep modes, load balancing, cell activation, beam management, and traffic routing to reduce unnecessary energy consumption in the infrastructure. The benefit comes from operating the network more intelligently under conventional traffic patterns, rather than from reducing the information content of transmissions;
- **Terminal EE**: Syntax-oriented AI can improve terminal EE through smarter receiver operation, adaptive discontinuous reception, power control, beam selection, and efficient

processing strategies. It can also reduce repeated transmissions and unnecessary channel measurements. The gains can be meaningful, although they depend on the complexity of the AI algorithms running on or near the terminal;

- **Mobility**: The impact on mobility can be strong and clearly positive. Syntax-oriented AI can enhance handover prediction, beam tracking, trajectory estimation, and multi-connectivity management, thereby maintaining service continuity at high speed. In this case, AI improves the robustness and efficiency of mobility procedures while preserving the conventional interpretation of the KPI;
- **Indoor and outdoor positioning accuracy:** The impact can be significant and positive. Syntax-oriented AI can improve positioning accuracy by processing conventional measurements such as Channel State Information (CSI), timing, angle, and signal strength more effectively, especially in multipath-rich indoor and dynamic outdoor environments.

3 6GARROW Use Cases Families & Use Cases (Revisited)

This section addresses the comprehensive taxonomy of use cases that are addressed throughout the project. The use cases are organized into three families:

- AI-native 6G design empowering new network services for stakeholders;
- AI-native 6G design for substantially more efficient & sustainable network operation;
- AI-native 6G design for autonomous networking.

In deliverable D2.1, each use case was documented with a detailed technical description accompanied by a business case rationale; an analysis of the relevant state of the art in the domain; an explanation of how the project's contributions extend beyond this state of the art; comprehensive functional and non-functional requirements and a set of KPIs and KVIs to measure success and impact.

This section is intended to refine those use cases based on the evolution of the state of the art, on new considerations about KVIs and KPIs (section 2) and on the feedback from the advisory board. Therefore, for each use case the structure is as follows:

- Summary of the use case, including technical description, business case and KPIs/KVIs;
- An update on the state of the art;
- New considerations;
- Identified risks, mainly from ethical perspective.

3.1 AI native 6G design empowering new network services for stakeholders

This use case family explores how AI-native 6G architecture empowers a new generation of intelligent services for both professional and commercial users. It includes AIaaS models that allow stakeholders to request customized AI model creation using obfuscated MNO data, enabling data monetization while preserving privacy. It also showcases ultra-reliable and low-latency machine vision-based robot control, such as intercontinental coffee robot operations, powered by GPT-based semantic communication. Furthermore, it addresses collaborative spectrum sharing through AI-powered radio measurement analysis. Collectively, these use cases illustrate how AI-native 6G architecture can support scalable, privacy-preserving, and context-aware services for both industrial and consumer domains.

3.1.1 AI-as-a-Service for stakeholders & creation of tailored AI models according to End Users needs

1 Summary

This use case enables users and other 6G stakeholders to request the creation or refinement of AI models leveraging both confidential or intra-network MNO data and proprietary deployments based on advanced AI techniques. This approach enables MNOs to monetize their datasets while maintaining data privacy through model-level obfuscation, preventing exposure of any sensitive underlying information.

1.1 Technical Description & Business Case

Diverse vertical market sectors are increasingly demanding personalized AI models on demand, also known as AIaaS, which leverage confidential and intra-network MNO data, such as the extensive

data repositories and behavioral patterns that MNOs accumulated through years of network operations and analysis.

Such comprehensive data collection would facilitate granular analysis of dynamic patterns, encompassing temporal variations in traffic flows across different time scales, i.e., hourly fluctuations, weekly cycles (weekends vs. weekdays), and seasonal patterns (school holidays in specific regions). This rich dataset could serve various market verticals, for example by enabling enhanced e-commerce personalization, targeted advertising strategies, and location-based promotional campaigns [Sundar_25], by supporting real-time decision making for navigation and traffic control of autonomous vehicles [Kim_24], innovative and environmental friendly novel applications in smart cities scenarios [Revathy_25], and coordination and deployment of rescue and recovery missions [Chumyen_25].

With Generative AI (GenAI) models and AI agents becoming increasingly integrated into operations across numerous market sectors, the combination of data-driven analytics and symbolic reasoning enables customized, context-sensitive decision-making capabilities across multiple industries, generating significant new business opportunities for both market leaders / incumbents and emerging Small and Medium Enterprises (SMEs) or innovative startups.

It is worth mentioning that the few ongoing trials currently based on advanced GenAI models and/or AI agents mostly rely on closed, proprietary solutions, oftentimes deployed side-by-side with incumbent/traditional (non AI-based) systems, to refine the performance and check the accuracy of the trained models. As commercial services based on native AI-networks compliant with 6G standards are not expected to be deployed before 2030[1], the market for solutions leveraging the AIaaS approach is still in its very early stages of strategic planning, ecosystem creation, and RoI evaluation [Dataeconomy_report_26].

The 6GARROW approach allows using MNO internal data to train tailored AI models according to the needs of individual users or groups of users. The proposed approach enables MNOs to monetize existing data while avoiding the exposure of the underlying raw data sets, as sketched in Figure 1.

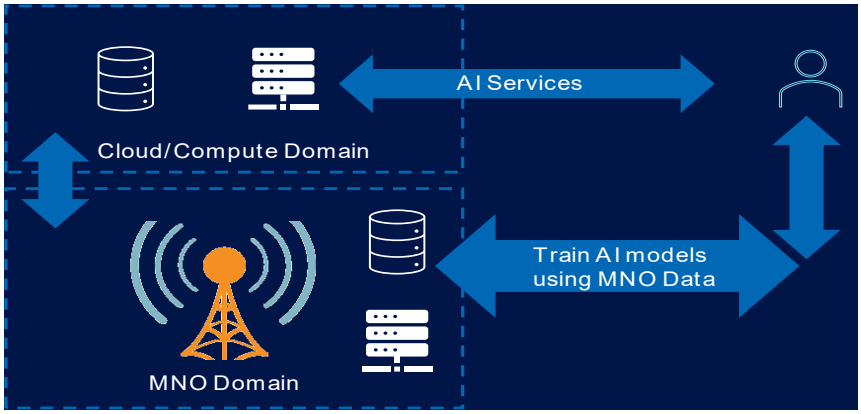

Figure 1: 6GARROW approach: MNO offers trained AI models exploiting MNO data, thus avoiding exposure of underlying raw training data.

1.2 KPIs and Requirements

1.2.1 Functional and Non-Functional Requirements (updated)

To enable the AIaaS vision, several functional entities need to be introduced, including:

[1] This expectation is based on the fact that the in the 3GPP calendar of work, a decision on the 3GPP Release 21 roadmap (the first normative 6G release) is expected at the June 2026 plenary meeting, and potentially later. This roadmap is expected to define a two-year timespan for completing the first 6G release, after which an additional 12 to 18 months would be needed for certification and initial deployments, which will then enable the launch of commercial services.

- Database containing MNO-internal data, not suitable for sharing externally in its raw form (however, sharing of processed results, for example an AI model derived from MNO-internal data is assumed to be suitable under certain conditions);
- Database containing non-confidential (public) data;
- Processing equipment for AI training (creation and refinement of models), mainly based on Graphics Processing Units (GPUs) [GPU_NVIDIA_26];
- Processing equipment for AI inferencing, mainly based on Central Processing Units (CPUs) [CPU_Intel_26];
- Novel architecture leveraging on both GPUs and CPUs [Blackridge_report_25];
- AI/Data Management entity receiving requests for Data/AI related processing and coordinating access to MNO-internal and non-confidential data and related processing equipment.

In addition, a number of interfaces need to be introduced, including:

- Interface enabling access to AI/Data Management entity to request creation and/or refinement of AI models;
- Interface connecting intra-Network entities to extra-Network (for example Cloud) entities for trusted processing of MNO-internal data on external (Cloud) entities.

1.2.2 KPIs and KVIs

This use case is mainly assessed by the following KVIs:

- Availability of new services to intra- and extra-network users and stakeholders to request the creation and/or refinement of AI models. An increased experience is provided through the usage of MNO-internal data for such refinement steps;
- Reduction of overall silicon needed for execution of AI processing through appropriate centralization of AI resources where possible and related optimization of overall network EE.

2 State of the Art

We did not identify new sota from the last deliverable (D2.1). However, some additional recent documents (references) have been added to the previous section.

3 New perspectives

No new perspective has been identified with respect to the previous deliverable (D2.1).

4 Ethical Risks

This use case focuses on preventing sensitive personal data from being exposed outside the MNO domain, thereby ensuring inherently more secure and privacy-preserving operations.

Furthermore, when assessed against the latest AI Act requirements using the European Commission's AI Act Explorer tool [AI_Act_Explorer_26], this use case:

- does not incur in any ***Prohibited AI practice***;
- cannot be considered creating a ***High-Risk AI system***
    - unless (depending on the implementation details) the MNO datasets and AI models are used for safety components in the management and operation of critical digital

infrastructure, road traffic, or in the supply of water, gas, heating or electricity (one of the mentioned-above market verticals potentially leveraging this use case).

3.1.2 Ultra-Reliable and Low Latency Communications for Machine Vision-based Robotic Control

1 Summary

Remote robotic control is becoming increasingly important for industrial automation, healthcare, smart environments, and geographically distributed cyber-physical systems [Y23]. In such applications, robots must interpret visual scenes, receive user command, and execute actions with high reliability and low latency. Conventional approaches often rely on transmitting raw images or video streams to support remote perception and control, which creates substantial bandwidth demands and may degrade real-time (RT) responsiveness, especially over long-distance networks. This motivates the development of semantic communication frameworks in which only task-relevant information is extracted, transmitted, and reconstructed for control [ABB+22].

1.1 Technical Description & Business Case

This use case explores intercontinental machine vision-based robotic control using a coffee robot as a representative example. The proposed robot, illustrated in Figure 2, employs a vision-language-action (VLA) model to extract higher-level semantic understanding from raw Red–Green–Blue–Depth (RGB-D) sensor data and execute text-based commands over an international network. This enables the robot to transmit only task-relevant semantic information, such as intent, status, or control signals, instead of raw image or video data. As a result, bandwidth consumption is reduced while responsiveness and control accuracy are improved. At the receiver side, the semantic information is reconstructed to support accurate remote control and task execution.

This use case is implemented as shown in Figure 2:

(a) The VLA model processes RGB-D sensor inputs to estimate the liquid level in the cup and updates the robot's internal state.

(b) Using the user command together with the inferred internal state, the VLA model generates control signals that guide robotic manipulation and task execution.

To support RT responsiveness, the system leverages Ultra-Reliable and Low Latency Communications (URLLC) as defined in 3GPP [22.261], which targets highly reliable transmission of small packets with extremely low latency. This capability is essential for precise actuation and continuous feedback in robotic control, especially when the system operates across long distances.

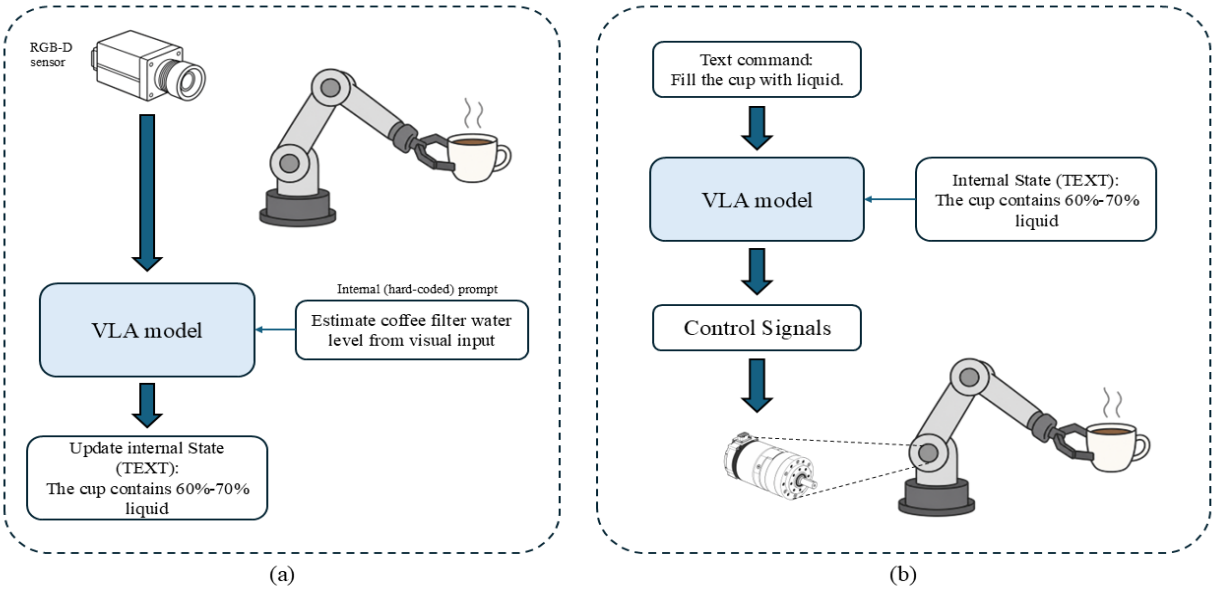

Figure 2: Intercontinental URLLC machine vision-based robotic control

Beyond coffee robots, the same framework can be extended to industrial robotics, pharmaceutical automation, and smart home applications, where low-latency and high-precision control are required [ABB+22], [OKP+24].

However, several challenges remain, including efficient large language model (LLM) deployment on edge devices through pruning [MFW23] and distillation [GDW+23], synchronization between distributed models [WZS+24], [LSZ+24], semantic-aware scheduling [DIN23], [ZC+23], and security and privacy protection [MPS21], [ZC22].

The business case for this use case is supported by three primary arguments:

- **Semantic cost reduction**: Semantic communications results in substantial reduction of bandwidth and infrastructure cost in remote robotic-control applications. By transmitting task-relevant semantic information rather than raw image or video streams, the proposed VLA-based system can reduce network load while preserving the information needed for accurate task execution;
- **Long-distance robotics**: A second value proposition is the ability to support mission-critical, low-latency and high-precision robotic services across long distances. This can enable new service offerings for MNOs and industrial service providers, such as industrial robotics, pharmaceutical automation, and smart home applications where precise actuation and continuous feedback are required;
- **Robot deployment in constrained or temporary scenarios**: The use case also creates business opportunities in constrained or temporary deployment scenarios where full infrastructure support is expensive or unavailable. By combining semantic communication with device–network collaboration, GenAI, and heterogeneous AI-agent communication, the system can support remote robotic operation in disaster recovery, large public events, and temporary industrial deployments. In these environments, the main economic advantage is not only lower data transmission cost, but also faster deployment, reduced operational complexity, and improved service continuity.

1.2 KPIs and Requirements

1.2.1 Functional and Non-Functional Requirements

The proposed system requires 6G networks to combine URLLC with semantic intelligence to support seamless intercontinental machine vision-based robotic control. Functionally, it needs semantic communication protocols, low-latency transmission interfaces, dynamic slicing, edge offloading, remote-control application programming interfaces (APIs), and compact edge-deployable AI models. Non-functionally, the system must satisfy strict latency, reliability, bandwidth-efficiency, EE, security, privacy, and robustness requirements so that robotic control remains accurate, safe, scalable, and dependable across long-distance deployments.

1.2.2 KPIs and KVIs

This use case is mainly assessed through reliability and U-plane latency. High reliability is required because similar robotic systems may be used in mission-critical industrial, cleanroom, or space-operation scenarios, where uninterrupted operation is essential for safety and task success. At the same time, very low U-plane latency is needed to enable RT interaction, near-instantaneous responsiveness, and precise actuation during remote robotic control.

2 State of the Art

As discussed earlier, conventional remote robotic-control systems often rely on raw image or video transmission, which leads to high bandwidth consumption and degraded RT performance. Recent studies address this limitation through semantic communication, where only task-relevant information is transmitted. Key advances include:

- **GeSa-XRF** [YXQ+24] introduces a generative semantic-aware extended reality (XR) framework, emphasizing *what to transmit* instead of *how to transmit*, enabling efficient multimodal communication;
- **GAI-SCN** [XSZ+23] proposes a cloud-edge-mobile semantic communication network supporting multimodal semantic provisioning and semantic-level joint source-channel coding;
- **Generative semantic communication systems** [LDS+24] demonstrate up to 99.98% reduction in communication overhead and significant accuracy improvements compared to traditional methods;
- **AI-enabled URLLC resource allocation** [Zha+24] applies deep reinforcement learning (DRL) and multi-layer optimization to achieve low-latency and reliable communication for Beyond 5G (B5G) / 6G;
- **Task-oriented semantic communication for URLLC** [Liu+23] integrates semantic importance into wireless control systems, improving delay performance and task execution accuracy.

Compared to existing work, 6GARROW extends semantic communication to intercontinental robotic control, where communication is driven by semantic intent rather than media transmission, enabling efficient, scalable, and RT global robotic orchestration.

3 New perspectives

Uncertainty-aware opportunistic Hybrid Language Model (U-HLM)

On the one hand, inference with VLA models requires substantial computational resources; on the other hand, the target application imposes stringent low-latency communication requirements. This creates a significant challenge for deployments in the UE, which is often constrained in terms of processing capability, memory capacity, and energy consumption. Executing a full VLA pipeline

locally on the User Equipment (UE) may therefore be impractical, particularly for mobile or lightweight devices.

To address these limitations, we employ a hybrid inference framework that distributes the inference process between the UE and an edge or cloud server [OKP+24]. In this scheme, shown in Figure 3, the UE performs an initial inference pass using a lightweight small language model (SLM), enabling fast local processing and reducing end-to-end (E2E) response latency. Although this local inference is computationally efficient, the reduced model size may lead to lower prediction accuracy and increased uncertainty for specific output tokens or action decisions.

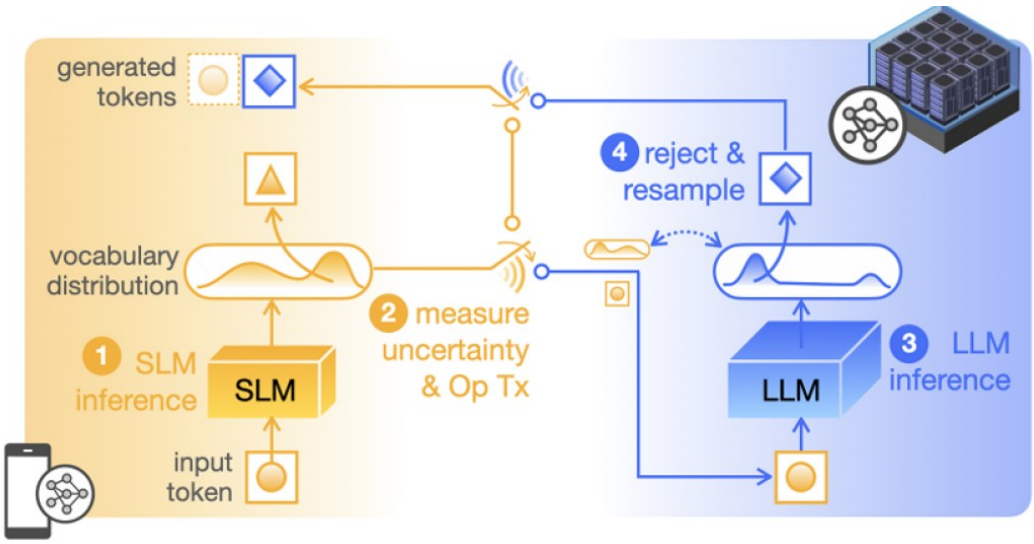

Figure 3: Schematic of the proposed U-HLM over a wireless network [OKP+24].

To maintain both responsiveness and inference quality, the SLM deployed on the UE continuously estimates the confidence of the generated tokens. Tokens associated with high uncertainty are selectively offloaded to a remote server hosting a more capable LLM. The LLM then performs verification and refinement of the uncertain outputs, correcting tokens when necessary and returning the updated results to the UE. By transmitting only uncertainty-critical tokens rather than the entire inference context, the proposed framework significantly reduces communication overhead while still benefiting from the superior reasoning capability of the LLM.

GPU-accelerated next generation node B (gNB) for semantic URLLC robotic control

In the proposed use case, semantic communication reduces the amount of transmitted data. However, the remaining packets are often highly task critical. Therefore, reliable and low-jitter delivery remains essential. A GPU-accelerated layer 1 (L1) can therefore accelerate compute-intensive baseband functions such as channel coding/decoding and other PHY operations [Nvi25b].

GPU-accelerated gNB for semantic URLLC robotic control is an architecture in which PHY processing is offloaded to the GPU, while higher-layer protocol functions remain on the CPU [Nvi25a]. It can not only reduce radio access network (RAN)-side processing latency, improve timing stability, and increase scalability under demanding wireless workloads, but also free CPU resources for AI inference, semantic processing, and Control-plane (C-plane) functions.

From a system perspective, this architecture enables a clean separation between wireless acceleration and application intelligence. Semantic encoding, semantic reconstruction, multimodal reasoning, and task orchestration can be handled by AI models at the edge or cloud, while the GPU-accelerated radio stack improves the responsiveness and robustness of the wireless link.

Integration with highly efficient communications with light infrastructure

Using selected studies from [6GARROW-D31] and [6GARROW-D41], the proposed use case can be extended beyond intercontinental robotic control toward broader semantic and goal-oriented communication scenarios. These potential extensions are outlined below.

- **Resilient operation under constrained network conditions**: Using semantic and goal-oriented data representation and adaptive RA mechanisms, the proposed use case can be employed in scenarios where bandwidth, energy, and connectivity are limited. This makes remote robotic-control more relevant to disaster recovery, military operations, large public events, and temporary deployments, where infrastructure may be weak, damaged, or costly to deploy;
- **Device-network collaborative framework**: The system can further evolve into a device–network collaborative framework using application-driven semantic negotiation and adaptive encoding. Terminals, edge nodes, and network infrastructure can jointly decide what semantic content should be transmitted, when it should be transmitted, and where semantic processing should take place;
- **GenAI integration**: GenAI can also be integrated directly into the communication pipeline. Instead of being used only as a post-processing module, generative models can support semantic compression, reconstruction, and interpretation under constrained bandwidth;
- **Heterogenous AI agent communication**: Finally, the use case can be generalized to heterogeneous AI agent communication, where robots, Unmanned Aerial Vehicles (UAVs), control systems, and human operators exchange task-level semantics rather than raw data. This requires interoperability across heterogeneous representation spaces. It also requires device–edge co-inference and energy-aware semantic extraction, where terminals balance local processing, transmission cost, and inference performance.

Human-in-the-loop and shared autonomy

In this use case, control is assumed to remain fully autonomous when the VLA model operates within its validated task domain and the semantic interpretation confidence is sufficiently high. This includes routine situations where sensor inputs are clear, the user command is unambiguous, the inferred internal state is consistent with the observed scene, and the planned action satisfies predefined safety and task constraints.

Human intervention becomes necessary when the system detects semantic uncertainty, conflicting interpretations, abnormal sensor inputs, or a mismatch between the command, the reconstructed semantic state, and the expected physical outcome. Examples include unclear liquid level estimation, occluded objects, unexpected object placement, ambiguous user intent, degraded communication quality, or low confidence in the generated control action.

Fallback interaction modes should, therefore, be defined to allow for human intervention. In low-risk situations, the robot may ask a clarification question or request an additional command from the user. In moderately risky situations, it may switch to a constrained autonomy mode, where only safe preparatory actions are allowed while awaiting confirmation. In high-risk or safety-critical situations, the system should enter a safe stop mode, preserve the current state, notify the operator, and transfer control to manual or supervised teleoperation until semantic interpretation is restored.

4 Ethical Risks

The proposed machine vision-based robotic-control system introduces ethical risks related to privacy, security, safety and impartiality.

- **Privacy and data protection:** The proposed remote robotic-control use case reduces privacy exposure by avoiding continuous transmission of raw RGB-D images or video.

Instead, only task-relevant semantic information, such as liquid level, object state, user intent, or control signals, is exchanged. However, these semantic representations can still reveal sensitive contextual information, including human presence, behaviour patterns, workspace layout, or operational routines. Therefore, semantic data should be treated as potentially sensitive and protected through data minimization, access control, and secure storage policies;

- **Security and robustness**: Security is also critical because incorrect or manipulated semantic information may directly affect robot behaviour. An attacker could attempt to modify semantic embeddings, inject false commands, alter uncertainty estimates, or interfere with the communication link between the UE, edge server, and cloud-hosted model. Such attacks could lead to unsafe actuation, incorrect task execution, or denial of service. The system should therefore include authenticated communication, integrity protection, encrypted transmission, continuous output validation, logging, and human-in-the-loop override mechanisms in safety-critical cases;
- **Explainability and accountability**: The limited explainability of VLA models can make it difficult to determine why a specific semantic interpretation or robot action was produced. This is especially important when decisions are distributed across the UE, edge server, and cloud-hosted model. The system should therefore support interpretable uncertainty estimates, traceable decision logs, confidence reporting, and mechanisms for human review when model outputs are uncertain, inconsistent, or safety-critical.
- **Fairness and generalization**: The VLA model may perform unevenly across different environments, object types, lighting conditions, accents, or user commands if the training and validation data are not representative. Such bias can lead to systematic errors in semantic interpretation or task execution. Evaluation should therefore include diverse visual conditions, user profiles, and operating scenarios.

3.1.3 Semantic Communications Services for Agents

1 Summary

AI agents are anticipated to become ubiquitous soon, evolving from simple request-response services for human users to autonomous entities capable of complex interactions. This evolution will drive extensive Machine-to-Machine (M2M) communications based on embedding representations, potentially creating significant network capacity challenges. These challenges arise because embedding-based data representations, though powerful, often include more data than is needed for specific types of communication.

1.1 Technical Description & Business Case

MNOs could provide differentiated semantic communication services for AI-agent traffic by negotiating Semantic Service Level Agreements (SSLAs) with applications through dedicated semantic interfaces. This would enable optimized handling of embedding-based exchanges, including adapted encoding, distortion monitoring, and transport mechanisms. Because AI-agent services are latency-sensitive, such services would rely on geographically distributed computing resources, including within the operator network, while also requiring efficient transfer of communication context to preserve continuity, especially in roaming situations. For MNOs, this creates both monetization opportunities and broader operational benefits, such as improved capacity usage, reduced Total Cost of Ownership (TCO), and better user experience. At the same time, it significantly increases their responsibilities regarding privacy, security, compliance verification, and accountability in increasingly autonomous and dynamic AI-agent ecosystems. Beyond user-facing services, AI agents could also enhance network operations through Operations, Administration, and Maintenance (OAM), intent-based networking, resource optimization, predictive maintenance,

anomaly detection, cyber threat prevention, and dynamic network slicing, reinforcing the operator's role in optimizing protocols and data representations for embedding-based communications.

1.2 KPIs and Requirements

1.2.1 Functional and Non-Functional Requirements

The system must deliver low-latency, real-time performance for both human-agent and agent-agent interactions, supported by a distributed and scalable computing infrastructure with efficient workload placement, including edge resources. It should provide reliable operation under varying load and network conditions, with strong fault tolerance, redundancy, failover, and graceful degradation mechanisms to preserve service continuity. The framework must also support semantic communications through suitable codecs, protocol adaptations, and Quality of Service (QoS) handling, while offering standardized, open, and developer-friendly interfaces that enable interoperability, integration, and efficient service development.

At the same time, the system must ensure strong security, privacy, and governance. Agent operations should be strictly controlled through robust authentication, authorization, and secure execution environments, while sensitive user and operator data must be protected through access control, integrity assurance, communication monitoring, and policy enforcement. Overall, the framework should promote standardization and interoperability across platforms and implementations, ensuring secure, flexible, and compliant support for advanced agent-based network services.

1.2.2 KPIs and KVIs

KPIs:

- U/C-Plane Latency: Today's bottleneck in LLM and agent usage primarily stems from inference computation time. As smaller, more efficient models are developed for smartphones and cost-sensitive applications, we can anticipate widespread deployment of agents across networks. These distributed agents will rely heavily on communication services to complete tasks efficiently. To support this ecosystem, edge servers must be strategically positioned throughout the network topology, with associated tools, contexts, and data served optimally to minimize response times and enhance user experience. Moreover, dedicated transport protocols and scheduling mechanisms should be developed to account for the specificities of AI traffic;
- Compression Rate: Closely tied to latency objectives, the semantic communications field aims to efficiently compress representations exchanged between agents. Effective compression algorithms must preserve task integrity and model interoperability while reducing transmission overhead. The ideal compression approach balances semantic fidelity with processing requirements, ultimately benefiting both system responsiveness and energy consumption, a critical consideration for sustainable AI infrastructure deployment at scale.

KVIs:

- User Experience Measurement: Behind agent communications lie the application connected to user requests or services. Semantic communications and AI-native 6G networks should configure equipment and services with application-specific metrics as primary considerations. This user-centric approach ensures that technical optimizations directly translate to tangible improvements in service quality, responsiveness, and overall satisfaction, creating a virtuous cycle of adoption and utilization;
- Return On Investment: Implementation of agent communications and dedicated services such as Compute-as-a-Service (CaaS) or AIaaS represents new monetization streams for operators while leveraging a natural evolution from software-based 5G architecture. These services should be introduced incrementally, allowing operators to amortize investments while gradually expanding capabilities and service offerings that capitalize on existing infrastructure investments;

- Trust: A significant concern surrounding AI and agents is the potential for data leakage, privacy violations and unintended actions. As service providers operating on their networks, operators bear responsibility for client data protection. This underscores the necessity for comprehensive security measures including data privacy safeguards, robust authentication mechanisms, and granular authorization controls. AI agents must be properly authenticated on the network, while the network itself must be capable of verifying legitimate AI agent services to maintain system integrity and user confidence.

2 State of the Art

The updated version of the 3GPP Technical Report (TR) 22.870 [22.870] significantly broadens the role of AI services and AI agents in the 6G service landscape. Compared to the previous version, which already considered some AI service and agent-related scenarios, the document now introduces a much larger set of AI-centred use cases, covering not only AI-assisted applications and distributed computing support, but also AI-agent-based interaction, service orchestration, network assurance, AI-native traffic handling, and AI-agent management. Altogether, the latest version of the TR confirms more clearly that 6G is expected to be AI-native.

The updated TR considerably expands the operational role of AI agents in the network and service environment. Beyond the use cases already present in the earlier version, the latest version adds multiple scenarios where AI agents act as explicit service and coordination entities. This includes 6.21 on “6G network providing on-demand networking with AI Agent”, 6.32 on “disaster rescue planning enabled by network AI Agents”, 6.43 on “AI Agent for network performance assurance”, 6.44 on “customised service provisioning based on AI Agents”, 6.45 on “flexible UE-Network coordination through AI Agent(s)”, 6.47 on “proactive AI Agent for personal safety”, 6.51 on “network-based intelligent assistance by a network-native AI Agent”, 6.55 on “shared embodied AI Agents”, and 6.62 on “AI Agent assisted rescue in the water park”. Compared to the previous version, this represents a major evolution because AI agents are increasingly treated not only as user-side assistants, but also as entities involved in service exposure, feasibility checking, resource coordination, safety support and network assurance.

Another important extension in the updated TR concerns trust, control and governance of AI agents and AI-enabled services. The introduction of use case 6.41 on “authentication and authorisation for AI Agents” makes explicit the need to treat AI agents as distinct actors whose actions and access rights need to be controlled. In addition, use case 6.46 on “AI Agent management” introduces network-level functions for the registration, discovery and invocation of third-party AI agents, while use case 6.56 on “feasible intent achievement for AI Agents” adds the important notion that the network should check whether an intent issued by an AI agent can actually be fulfilled before resources are committed. These additions were not present in the earlier version and are particularly relevant because they show that 3GPP is moving beyond simple AI enablement towards a more operational treatment of AI-agent trustworthiness, accountability and controllability.

Huawei’s A2A-T initiative provides useful context for this discussion on agent-to-agent communications in telecom environments. Within the TM Forum, “Agent-to-Agent for Telecom” (A2A-T) is being positioned as a telecom-oriented interaction framework for multi-agent collaboration, especially in autonomous network operations [IG1453], [IG1453A]. At its core, A2A-T provides a unified framework for AI agents to interact, collaborate, and orchestrate tasks across different network layers and domains. The proposal should be seen as a specialization of the more common A2A framework from Google / the Linux Foundation. In this context, A2A-T appears as a connected but separated initiative, rising practical questions regarding the maintainability and synchronisation with the official A2A protocol.

Finally, a particularly important addition in the TR 22.870 is the use case 6.59 on “6G provide communication service for AI traffic”. This use case is especially relevant because it addresses AI traffic explicitly as a communication subject, rather than only as a by-product of AI applications. More specifically, it recognizes that AI applications increasingly exchange information in forms such as

tokens, and that this traffic may exhibit communication properties that differ from conventional application traffic. The use case highlights that tokenization can significantly reduce the amount of data to be transmitted compared to raw multimodal data, while still preserving information useful to the AI task. It also points out that different tokens may not have the same importance, since some contribute more directly than others to the quality of inference or generation. In addition, the use case considers that AI traffic may tolerate a certain level of errors, because some missing or corrupted tokens may be inferred, reconstructed or compensated for by the receiving model. This is a particularly important shift compared to conventional communication assumptions, where bit-level fidelity is generally treated as uniformly critical. The use case also stresses that AI traffic may have specific temporal properties. Depending on the application, token exchanges may be generated in burst form, in streaming form, or in close interaction with AI inference loops. As a result, latency constraints are not only linked to packet transport itself, but also to the timing of token generation, token processing, and overall interaction between communicating AI systems. The text further distinguishes different communication contexts, such as human-to-agent, robot-to-agent and agent-to-agent communications, each of which may impose different requirements in terms of latency, reliability and tolerance to degradation. Although TR 22.870 does not yet formulate this explicitly in terms of semantic communication or embedding-native transport, use case 6.59 is particularly significant because it is one of the clearest 3GPP acknowledgements so far that AI-oriented traffic may require differentiated communication treatment depending on the structure, role and relative importance of the information being exchanged.

More generally, current 3GPP discussions indicate a growing awareness that AI traffic cannot be efficiently handled using only the assumptions and mechanisms inherited from conventional mobile broadband services. Ongoing contributions highlight that AI-generated and AI-consumed traffic may be highly bursty, phase-dependent, asymmetric between uplink and downlink, and subject to rapidly changing latency and reliability constraints within the same application session. In particular, several discussions point to the need for finer-grained QoS handling, with more dynamic support for differentiated treatment inside what would conventionally be seen as a single service flow. This reflects the broader understanding that conversational, GenAI and agentic AI traffic may require more flexible and adaptive QoS mechanisms, although discussions remain at a general level and no stable conclusion can yet be drawn.

3 New perspectives

No new perspective has been identified with respect to the previous deliverable (D2.1).

4 Ethical Risks

Ethical deployment of AI agents in telecom and networked environments depends fundamentally on security, trust, and accountability. Agents are not only exchanging information; they may observe, decide, and trigger actions across services, APIs, and infrastructure. This means that ethical concerns such as user protection, privacy, safety, and control cannot be separated from the mechanisms used to authenticate, authorize, and monitor agent behaviour. A key lesson from the STIR/SHAKEN experience is that proving who initiated a communication does not by itself prove that the requested action is legitimate [STIR/SHAKEN]. Identity-based trust is therefore necessary but insufficient for agent operations, especially when agents can invoke tools, call other agents, and propagate decisions across multiple domains. In this context, ethics requires a broader trust model that considers not only identity, but also declared intent, execution authority, traceability, and post-action accountability.

In multi-agent environments, these risks are amplified because compromised or poorly governed agents can spread harmful actions across chained interactions. This is why recent telecom work increasingly aligns with zero-trust principles, as illustrated by the TM Forum “Zero Trust Agents for Autonomous Networks” project’s architecture, where no agent, tool, or interaction is trusted solely on the basis of valid credentials [C26.0.933]. Instead, trust must be continuously supported by pre-

runtime governance through agent registries, verifiable identity chains, policy enforcement at the agent gateway, controlled delegation, and tamper-evident observability across A2A, agent-to-tool, and agent-to-model interactions. From an ethical perspective, this approach is important because it helps preserve accountability and ensure that harmful actions can be detected and stopped before they propagate across services or network domains [C26.0.933].

3.2 AI native 6G design for substantially more efficient & sustainable network operation

The use case family "AI Native 6G Design for Substantially More Efficient & Sustainable Network Operation" focuses on integrating AI into the core architecture of 6G networks to enhance their efficiency and sustainability. This approach leverages AI to dynamically manage and optimize network resources, ensuring seamless connectivity and high-quality service for end users. By continuously monitoring network traffic, user demands, and environmental conditions, AI systems can allocate bandwidth, predict maintenance needs, and implement energy-saving strategies. This results in reduced operational costs and a lower environmental impact.

3.2.1 Highly efficient communications with light infrastructure

1 Summary

1.1 Technical Description & Business Case

In this use case, we rely on semantic-empowered RAN to guarantee reliable and energy-efficient communications even with light infrastructure (light infrastructure refers to deployment environments where communication and computing resources are limited, temporary, sparse, energy-constrained, or partially unavailable, requiring the network to operate efficiently with minimal support infrastructure). An AI-native semantic-empowered 6G network minimizes the amount of data exchanged, therefore allowing reliable audio, images and video exchanges and high autonomy of devices like UAVs and robots, with minimal energy and bandwidth usage.

The scenarios include disaster situations (storms, earthquakes), military operations, and events like music festivals and sports (marathons, ...). Reliable and energy-efficient communications are needed, including audio, image, and video services. The use case relies on coordinated UAVs, robots, sensors, and human operators.

Market opportunity and addressable verticals

The use case addresses three high-value market segments. Disaster response and emergency management is a global priority. Military and defense represent one of the largest and most stable purchasing budgets worldwide. The events sector is a commercially scalable market with recurring demand and clear willingness to pay for temporary network densification.

Core competitive differentiator

The concept of “lightweight infrastructure” is the main selling point. In all three scenarios, deploying conventional dense infrastructure is either impossible (disaster zones with damaged towers), prohibited (military operations), or impractical (temporary events). A semantic-native approach that does more with less bandwidth directly addresses a problem that 5G and conventional solutions cannot easily solve.

Revenue and business model angles

Network operators and defense contractors could license or embed the semantic processing stack into existing platforms. Event infrastructure providers could offer it as a managed service, charging per-event or per-seat.

1.2 KPIs and Requirements

1.2.1 KPIs and KVIs

For this use case, relevant KPIs – from section 2.2.1 – include:

- **Reliability**. Some services from this use case may have task completion for objective, e.g., automatic video-based recognition of victims in a disaster scene or identification of threats in a military operation based on UAVs observations. By transmitting only the most relevant information, the system may remain operational even under unfavorable channel conditions;
- **Terminal EE**. When it comes to UAVs or robots, EE is a crucial parameter, allowing longer duration and higher range of action. Semantic communications lead to less raw data transmission, reducing required energy. On the other hand, it must be guaranteed that local AI inference does not cancel this positive benefit. Split inference or quantized ML may therefore be considered;
- **Bandwidth efficiency**. Semantic-oriented AI can significantly reduce the spectrum required for a given terminal, which may be of interest to operate more terminals or to lower the costs.

1.2.2 Functional and Non-Functional Requirements

Main functional requirements:

- Semantic extraction and encoding of task-relevant information;
- Support for heterogeneous devices and agent coordination;
- Context-aware QoS and communication-computation optimization.

Main non-functional requirements:

- Reliability under weak or damaged networks;
- Low-power operation and reduced bandwidth usage;
- Scalable support for distributed devices and agents;
- Robustness in dynamic environments.

2 State of the Art

This use case shows several similarities, in intention, with some use cases from a recent (Q1 2026) version of 3GPP TR 22.870:

- “Use case on optimizing 6G infrastructure utilisation via resource exposure in 6G” proposes to leverage computational and storage infrastructures available at the edge of the network;
- “Use case on intelligent UAV swarms” uses UAVs swarms to collect data in case of public safety scenarios;
- “Use case on AI-based video analysis” delegates the analysis of videos to AI;
- Similarly, “Use case on network-assisted video-based AI inference task offloading for mobile embodied AI” processes and analyses video, but directly on mobile devices;
- “Use case on disaster rescue planning enabled by network AI Agents” deploys robots for rescue in a disaster scene, leveraging AI-agents for coordination;
- “Use case on ubiquitous emergency rescue via UAVs” utilises high-resolution-camera-UAVs and NTN on disaster scene;
- “Use case on cooperating mobile robots” enables a group of robots to cooperate and autonomously solve tasks.

Contrary to the 3GPP use cases presented above, in which semantic communications are not mentioned, this use case leverages semantic and goal-oriented communications for:

- Joint semantic intelligence at device and network levels;
- Enhanced collaboration among heterogeneous devices and AI agents;
- Dynamic resource allocation based on semantic importance and context;
- Task-oriented transmission instead of raw data exchange.

The SNS phase 2 project 6G-GOALS [6GGOALS] focuses on integrating semantic and goal-oriented communications within 6G. As such, it has defined several use cases which features may be related to the use case described in this section, such as:

- "Teleoperation" allows human operators to manipulate robots, using joysticks, haptic devices or Virtual Reality systems;
- "Collaborative robotics" involves several robots and human to collaborate to a common goal;
- "Semantic-enabled edge intelligence" performs data analysis and ML model execution directly on devices rather than in centralised cloud servers.

6G-GOALS shows that semantic and goal-oriented communications can be used to control robots (on a disaster scene) or to run AI models directly on devices. 6GARROW use case goes beyond the state of the art by integrating the above-mentioned semantic communications technologies into a 'disaster / temporary event' scenario. In this scenario, even with low energy available, the network can operate robots, swarms of UAVs, and provide multimedia information to operators.

3 New perspectives

No new perspective has been identified with respect to the previous deliverable (D2.1).

4 Ethical Risks

Semantic communications is a quite recent field of research, raising several questions on risks when relying on it, especially for 'disaster' scenarios and military operations where human life may be at stake. Identified risks are:

- Semantic misalignment between the sender and the receiver, possibly leading to wrong actions;
- Context dependency: if the context changes (i.e., the environment or task), incorrect interpretations may occur;
- Reduced traceability: as the access to raw data is limited, it becomes difficult to verify decisions, investigate incidents, or comply with regulatory requirements;
- Security: Attackers may target the semantic representation itself, which may be harder to detect with traditional security tools.

3.2.2 Dynamic resource allocation for intercontinental RAN-Core

1 Summary

With the rapid growth in mobile data traffic and diverse application requirements, such as video streaming and IoT-related services, static resource allocation leads to inefficiencies and degraded QoS [LLZ+21]. Network operators face challenges in dynamically managing resources across geographically distant locations, detecting and proactively mitigating anomalies. These challenges motivate the need for an intelligent, adaptive framework that can coordinate resources across regions while responding proactively to changing network conditions.

This use case enables AI-driven dynamic resource allocation (DRA) across two continents leveraging time-zone differences to optimize bandwidth and computing resources. This framework combines DRL-based decision-making, convolutional neural network (CNN)-based pattern extraction, recurrent neural network (RNN)-based demand forecasting, variational autoencoder (VAE)-based scenario generation, and autoencoder-based anomaly detection to optimize traffic steering, QoS prioritization, slice allocation, user plane function (UPF) selection, routing, and congestion mitigation. Unlike existing methods focused on isolated networks, this approach balances RAN and core network (CN) resources across regions, enabling scalable, efficient, and proactive 6G network management.

1.1 Technical Description & Business Case

The proposed DRA system, illustrated in Figure 4, connects a gNB in Korea with a core management core (CMC) in Finland through a virtual private network (VPN)-secured internet connection. The system integrates advanced ML models, including:

1. **Central controllers**: DRL agents act as central controllers, analysing network data from both the gNB and the CMC. DRL agents such deep deterministic policy gradient (DDPG), twin-delayed deep deterministic (TD3), soft actor-critic (SAC) [SNH+25], and proximal policy optimization (PPO) [BGV+22] can analyse information collected from network functions (NFs) such as the CMC Controller, UPF, Session Management Function (SMF), Access and Mobility Management Function (AMF), and policy-related entities. They monitor network indicators such as E2E traffic demand, flow-level congestion, packet delay, jitter, packet loss, bandwidth utilization, slice resource usage, QoS flow performance, and SLA violations [MDC+24]. Based on this network-wide view, they can dynamically optimize CN behaviour, including traffic steering, QoS flow prioritization, network slice resource allocation, UPF selection, routing path adjustment, congestion mitigation, and policy enforcement;
2. **Spatial pattern analysis**: CNNs could analyse spatial patterns across cells, slices, links, or time windows, and detect congestion regions. They could also be used as feature extractors before the DRL controller, summarizing complex network measurements into simpler patterns that the DRL agents can then use;
3. **Demand forecasting**: Network traffic often exhibits predictable spatial and temporal patterns that can be exploited by ML models. Short-term forecasting requires lightweight, low-latency models for rapid operational decisions, while long-term forecasting can rely on more expressive models for capacity planning. RNN models such as long short-term memory (LSTM), gated recurrent unit (GRU), as well as temporal CNNs or temporal fusion transformers (TFTs) are suitable choices for this application [KBZ+24], [ZGJ+19], [LAL+21];
4. **Traffic scenario generation**: In CN optimization, relying on a single traffic forecast may be insufficient because future demand can vary significantly under different user, service, and regional conditions. For such an application a VAE can generate multiple possible demand scenarios, allowing the orchestrator to test how the network would behave under normal load, peak-hour congestion, sudden traffic bursts, or regional demand shifts [KTD+25]. This is useful for proactive capacity planning, slice resource preparation, UPF selection, and congestion avoidance before the actual traffic arrives;
5. **Anomaly detection**: Abnormal behaviour in traffic, latency, packet loss, or resource utilization can indicate congestion, service degradation, network failures, or security threats. Because many abnormal events are difficult to define explicitly in advance, the system needs a mechanism that can learn normal network behaviour and detect meaningful deviations from it. Autoencoders or graph autoencoders (GAEs) are well suited for this purpose because they learn compact representations of normal traffic and QoS patterns and flag network states that cannot be reconstructed accurately [SLS+25]. This enables early detection of unusual network conditions before they escalate into severe QoS violations or service failures.

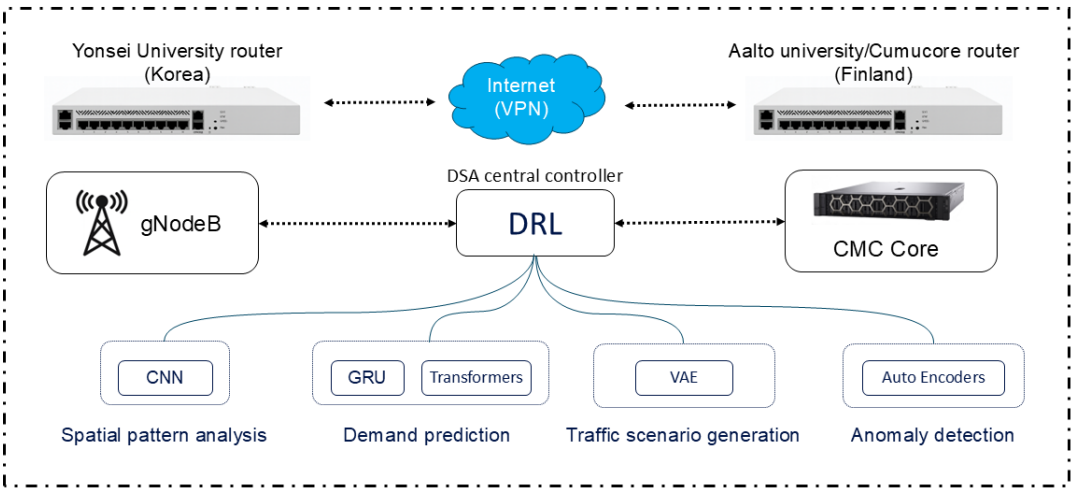

Figure 4: AI-driven intercontinental RAN-Core resource management architecture enabling dynamic and efficient coordination between gNB in Korea and CN in Finland

The DRL agents continuously monitor network performance and, based on predictions from LSTM/GRU or TFT models, dynamically allocate bandwidth, processing power, and other CN resources. CNNs can further support this process by extracting spatial and temporal traffic patterns across cells, slices, links, or time windows and providing simplified network features to the DRL agents. Generated traffic scenarios further allow the system to evaluate multiple possible demand conditions in advance, improving its ability to prepare for peak loads, sudden traffic bursts, and regional demand shifts. Finally, anomaly detection triggers alerts and resource adjustments to maintain optimal performance. All actions are logged and auditable so that operators can trace why a specific AI model made a certain decision.

Economic value and ROI validation

One of the main goals of the proposed architecture is creating business value. AI-based prediction reduces unnecessary overprovisioning and improves the utilization of compute, transport, and CN resources. Automation lowers operational costs by reducing manual intervention and improving EE, while better sharing of cloud infrastructure can delay expensive hardware expansion. The financial evaluation compares the AI-driven system against a traditional static deployment and also includes the cost of building and maintaining the AI infrastructure itself. Energy use and carbon-aware operation can also be included as additional evaluation criteria [JS21].

1.2 KPIs and Requirements

1.2.1 KPIs and KVIs

Table 4 shows a summary of the KPIs and KVIs in this use case.

Table 4: A list of KPIs and KVIs of use case DRA for international RAN-Core

KPI/KVI	Target or reported value
E2E Latency	eMBB target: ≤ 100 ms; measured median latency: 80 ms
Cross-Border Throughput	eMBB peak target: 20 Gbps; demonstrated average: 15 Gbps
Resource Utilization (Compute & Network)	CPU/spectrum utilization: > 70 %; improvement over isolated management: 20 %
QoS Compliance Rate	SLA targets: latency ≤ 50 ms, throughput ≥ 100 Mbps, packet loss ≤ 10^{-3}; achieved SLA compliance: 99.5 %; improvement over rule-based control: 15 %
Cost Savings	CAPEX/OPEX reduction: 30–45 %
EE Gains	Energy consumption reduction: 25 %
Scalability & Flexibility	Linear scaling with traffic and users while maintaining SLA targets

1.2.2 Functional and Non-Functional Requirements

The system requires a cross-domain AI orchestration framework with the following functional capabilities: AI-based decision-making for traffic steering, QoS prioritization, slice resource allocation, UPF selection, routing, and congestion mitigation; demand forecasting for short- and long-term resource planning; traffic-scenario generation for proactive capacity preparation; anomaly detection for abnormal traffic, latency, packet-loss, and resource-utilization patterns; and logging and auditing for resource-allocation decisions, model outputs, anomaly events, and rollback actions.

The system must provide predictable latency despite long-distance links and maintain service-level guarantees (SLGs). It should be able to scale to additional gNB-core connections without major redesign, while optimizing bandwidth, compute, and energy use to avoid waste. Because the use case is cross-border, it must respect data-routing laws, operator data sovereignty, and secure cooperation between involved operators. Finally, the system must remain robust during temporary network disruptions, ensuring that essential functions and logs continue to operate reliably.

2 State of the Art

As mentioned earlier, reinforcement-learning (RL) is central to our controller and has been extensively studied for resource allocation. For example [WLW+23] uses RL in RAN intelligent controller (RIC) to decide whether requests should be processed by the near-real-time (near-RT) RIC or by the non-real-time (non-RT) RIC. It has additionally been applied to RT DRA under changing traffic and device conditions in [LLZ+21], as well as to dynamic RAN and MEC resource allocation in Open-RAN (O-RAN) [MDC+24], [QWZ+25], and [SNH+25].

While most existing literature focuses on RAN–UE DRA, [KS21] models an E2E MNO architecture that includes both RAN and CN functions distributed across local, regional, and core data centres.

Forecasting is used to make the system proactive instead of purely reactive. Probabilistic forecasting has already been integrated into cloud-native RAN resource management through monitoring, analytics, decision, and actuation functions inside rApps [KBZ+24]. Other studies show that VAE-based models can represent uncertainty in demand prediction [KTD+25], while TFT models provide interpretable multi-horizon forecasting [LAL+21].

Safety and security are also relevant to this use case. Digital twins (DTs) can test and evaluate RAN behaviour before deployment [NST+25], [JS21]. Counterfactual KPI analysis estimates the likely outcome of alternative applications or policies [HPZ+25]. Finally, in [TMK+24], an AI/ML-driven intrusion detection system detects suspicious traffic in a real RAN testbed and triggers resource mitigation.

3 New perspectives

Scope and limits of international resource coordination

Not every network resource can be coordinated between Korea and Finland in the same way. Some resources must remain local because intercontinental delay exceeds their control-loop requirements. Others are inherently tied to local infrastructure and cannot be moved or shared across countries. Specifically, the following resources are immovable:

- **Resources limited by timescale**: RT and near-RT radio decisions must be handled locally because they require very fast reactions to changing radio conditions. Functions such as scheduling, link adaptation, fast congestion response, and immediate radio-layer reliability control cannot wait for intercontinental signalling [WLW+23]. Regional controllers can manage medium-speed actions such as traffic steering and slice adaptation, while the global orchestrator should focus on slower tasks such as long-term prediction, planning, and optimization across Korea and Finland;
- **Resources limited by physical nature**: Some resources cannot be coordinated globally because they are physically bound to a specific location. Radio resources such as spectrum, physical resource blocks (PRBs), antenna configuration, and local gNB scheduling capacity remain tied to the Korean radio site and cannot be transferred to Finland. In contrast, cloud resources, CN functions, compute capacity, transport policies, and network slices can be coordinated across regions.

Security, privacy, and trustworthiness

The intercontinental DRA system must protect both network operation and user-related data. Since the framework continuously monitors traffic parameters, it must minimize exposure of sensitive information and respect data-routing laws, operator data sovereignty, and cross-border privacy requirements. Security mechanisms are also needed to protect the VPN-based connection, prevent misuse of orchestration decisions, and ensure that anomaly detection does not trigger harmful or unnecessary mitigation. Malicious users may send excessive traffic that can potentially cause failures [TMK+24]. To address these attacks, our proposed autoencoder anomaly detector continuously monitors network traffic to classify users as normal or malicious and dynamically adjusts network resources in real time. Suspicious users receive fewer radio resources, while severe attacks such as Denial of Service (DoS) are mitigated. Trustworthiness should be supported through transparent logging, auditability, rollback mechanisms, and clear operator oversight, so that decisions made by AI components can be traced, reviewed, and corrected when needed.

4 Ethical Risks

Privacy and cross-border data exposure

The system depends on continuous monitoring of traffic, QoS metrics, latency data, anomaly signals, and resource-allocation logs between Korea and Finland. Even if the project uses decentralized learning, network telemetry can still reveal sensitive patterns such as mobility behaviour, application usage, device identifiers, or operational details of the network. The ethics document explicitly warns that personal data such as mobility patterns, application usage, IMEI numbers, or IP addresses require general data protection regulation (GDPR)-level protection, legal basis, transparency, anonymization or pseudonymization, and safeguards for transfer outside the European Union (EU) / European Economic Area (EEA). This is especially important here because the use case is explicitly cross-border and depends on international coordination between RAN and core domains.

Unfair or harmful resource prioritization

The AI controller reallocates bandwidth, processing power, and other resources based on predicted demand and anomaly detection. This creates an ethical risk that some users, regions, applications,

or service classes may be systematically deprioritized if the model optimizes only global efficiency or operator cost. For example, a low-revenue service, a smaller regional network, or unusual traffic pattern could be treated as less important even when the service is socially important. Since the ethics framework emphasizes fairness, accountability, transparency, robustness, and privacy, the resource-allocation logic must be evaluated not only by throughput or utilization, but also by whether it produces unfair service degradation across users and regions.

Accountability and safety in automated decisions

This use case gives the AI system a strong operational role: DRL-based monitoring, CNN spatial feature extraction, RNN/TFT forecasting, autoencoder-based anomaly detection, and automatic resource adjustments. If the AI model makes a wrong prediction or falsely flags traffic as anomalous, it could reduce bandwidth, trigger unnecessary mitigation, degrade QoS, or affect network stability. The ethics document mentions the need for logging and auditing of allocation decisions, inference outputs, and anomaly events, which is ethically important because operators must be able to explain, review, and reverse automated decisions. Without clear accountability, a harmful network action could be blamed on the model without a responsible human or organizational process behind it.

3.2.3 Overhead reduction through Semantic Closed Control Loop for Private 5G Management Systems

1 Summary

The Semantic State Representation Function (SSRF) is introduced in this use case as a new mechanism that enables more effective supervision and control of 5G and upcoming 6G networks. It transforms the way operational information is interpreted. As the size and complexity of modern mobile systems increase, the mix of Graphical User Interface (GUI)-based tools and isolated AI modules becomes insufficient to understand the global network condition. Each component generates independent traces, logs, and metrics that do not naturally combine into a meaningful picture. The SSRF overcomes this limitation by relying on advanced processing techniques and LLMs to extract semantic meaning from heterogeneous data streams across all network layers. It produces real-time representations that can be immediately understood both by humans and by automated agents. This enhances situational awareness for operators, accelerates decision-making, and enables scalable M2M interactions in which intelligent components share contextualized knowledge rather than raw, disconnected signals.

1.1 Technical Description & Business Case

Modern 5G Core (5GC) environments such as the HPE Aruba Private Networking 5GC [HPE] are becoming increasingly complex due to service based and cloud native architectures [MKP23]. NFs such as AMF, SMF, UPF, and PCF continuously generate large volumes of heterogeneous telemetry including logs, KPIs, counters, alarms, and performance traces. However, this information remains fragmented across multiple layers and tools, preventing a unified understanding of the network state. As a result, troubleshooting relies on manual correlation of distributed signals, which does not scale with the growing complexity of 5G systems. Intent Based Networking (IBN) and Intent Driven Management frameworks [3GPP+23], [XS25] define high level objectives, but current implementations still lack a consistent way to translate operational data into actionable insight [ISP25].

The SSRF (see Figure 5) addresses this gap by introducing a semantic layer across the 5GC that aggregates and interprets multi-domain telemetry. By leveraging advanced analytics and LLMs [PMS+22], SSRF transforms fragmented operational signals into coherent representations of network state, enabling a more unified and interpretable view of system behaviour and supporting intent driven decision processes.

From a business perspective, this approach reduces operational complexity and manual effort in troubleshooting, shortens time to resolution, and enables more scalable automation of network management, improving overall efficiency and reliability of 5G operations.

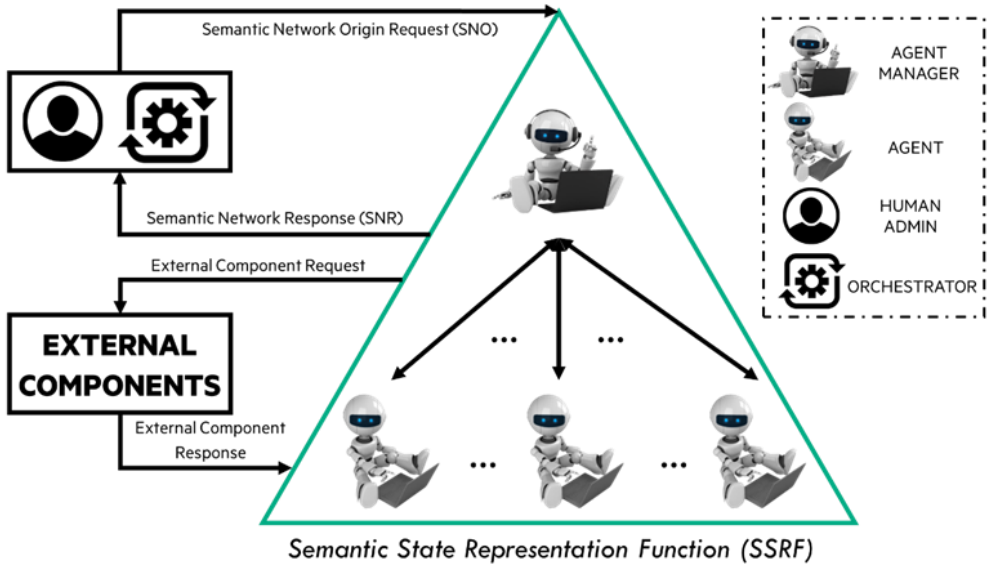

Figure 5:High-level SSRF visual architecture

1.2 KPIs and Requirements

1.2.1 KPIs and KVIs

This use case aims to show how using the SSRF can improve 5G Network Management by improving overall efficiency, reliability, and automation. This is accomplished by the SSRF generating semantic representations of heterogeneous network telemetry in such a way that they represent and reflect an accurate and meaningful representation of the situation, thereby enhancing situational awareness and enabling data-driven decision-making through KPIs, as shown below, and the corresponding expected performance and effectiveness of the project to provide interpretable and reliable network state information.

Table 5: Semantic State Representation Function KPIs

KPI/KVI	Target or reported value
RMSE Semantic State Accuracy	< 0.1
Semantic Similarity (SSM)	> 0.9
Answer Correctness (AC)	> 0.85

1.2.2 Functional and Non-Functional Requirements

No new functional or non-functional requirements have been identified with respect to the previous deliverable (D2.1), as the proposed solution remains consistent with the previously defined system specifications.

2 State of the Art

Mobile networks are getting more complex so managing them needs to be smarter and more automatic. IBN lets operators say what they want in simple words and the system changes to

implement the technical steps needed, which is a complex task because the system must understand what the operator means, make sure it works when things change, and coordinate different parts of the network. Lately, LLMs have been introduced to help solve the problem. LLMs work with tools like multi-level resolvers [LSX+25], semantic routers [DAA25] and systems that check predictions before acting [MPY+25]. Others have made new ways to measure how good the system's answers are like the FEACI score [LSX+25]. But still several problems remain. It is hard to keep these systems running in fulfilment of the given requirements when the network changes fast and many open-source models can be tough to set up or change. Another challenge is making sure the system's decisions are clear and flexible enough for real networks.

Today many network management systems use ML or DL for specific tasks like spotting faults or guessing traffic. However these tools can only handle part of the network. They do not give a full picture of what is happening. Also, although LLMs have helped humans and machines work better together in many areas, network management has not fully used them yet [AAC24]. With 6G networks coming, there is an increasing demand for systems that can bring together data from different sources and give operators a clear real time view of the whole network.

This shows the main weakness of current AI based network management. Most tools fix small problems instead of looking at the whole network. They might find faults or predict traffic well, but they do not usually manage to give operators a full up to date view. LLMs could help by understanding what operators want, combining data from different places and helping automate tasks. But to work well, LLMs need to be better linked with current management systems and integrate seamlessly with ML and DL tools.

3 New perspectives

No new perspectives have been identified with respect to the previous deliverable (D2.1).

4 Ethical Risks

The emergence of the SSRF involves certain ethical and social aspects that should be considered. Firstly, since the SSRF handles large amounts of data, which includes both logs and telemetry data produced by various NFs, problems with personal data privacy and security are expected to occur. This can be avoided by implementing appropriate data management strategies, using anonymized data, and complying with the data protection requirements.

The other type of ethical risk relates to the problems with biased and erroneous interpretation of information using AI methods in the SSRF. Using LLMs may lead to distorted or oversimplified results if they are poorly trained and tested for accuracy and reliability. This could cause improper action or mistrust towards AI-based solutions. To address this problem, the transparency of AI-based results and the ability to review them manually should be provided.

Another risk lies in possible misuse when such highly autonomous technology will be used for purposes not covered by the operational policy. Again, well-developed access management and audits should address this problem efficiently.

Overall, by focusing on explainability, human oversight, and responsible data handling, the SSRF can minimize ethical risks while providing tangible benefits to network operation and society.

3.2.4 RAN/UE & RAN/CN Cross-Domain Coordination

1 Summary

6G is expected to evolve toward an AI-native architecture where AI/ML is deeply embedded across the network. PHY AI is expected to improve channel estimation, CSI feedback, beamforming,

modulation, and receiver processing, while RAN-CN coordination to enable context-aware mobility prediction, resource allocation, traffic steering, and energy optimization.

1.1 Technical Description & Business Case

6G is expected to adopt an AI-native architecture in which AI/ML technologies are deeply integrated into the network to enable intelligent automation, performance optimization, sustainability, and resilience. As conventional wireless communication approaches its performance limits through isolated optimization of system components, AI/ML emerges as a key enabler for improving link performance, transmission efficiency, and adaptability in complex real-world environments.

Physical Layer AI/ML Technologies

AI/ML technologies improve resource efficiency and reduce energy consumption through optimized channel estimation, CSI feedback, beamforming, beam management, modulation, and receiver processing, particularly in massive Multiple Input Multiple Output (MIMO) and high-frequency systems.

RAN/CN Coordination

AI-native 6G systems require continuous exchange of intelligence between RAN and CN domains. RAN insights such as radio conditions and mobility can optimize CN functions, while CN information such as service requirements and user profiles can enhance RAN scheduling and resource allocation. This bidirectional coordination enables adaptive, context-aware, and E2E optimized network operations, as shown in Figure 6.

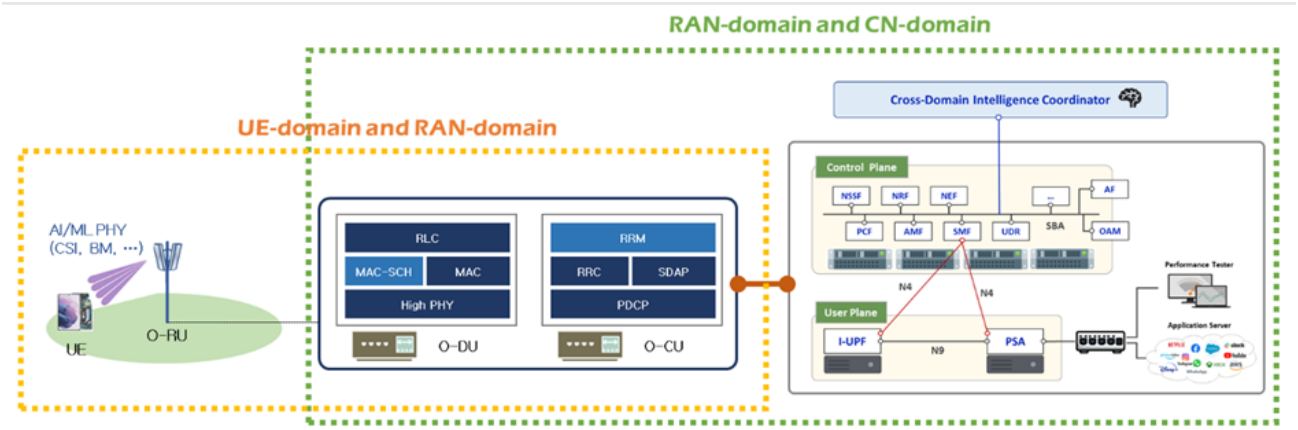

Figure 6: RAN/UE & RAN/CN cross-domain coordination

1.2 KPIs and Requirements

The key metrics to assess AI/ML-based RAN/UE & RAN/CN Cross-Domain Coordination technology include:

- **Latency**: The inferencing time shall not exceed the value defined in 3GPP specification;
- **Mobility**: performance gain of the cross-domain RAN-CN mobility prediction framework over single-domain baselines, where the baselines correspond to mobility prediction models trained and inferred using only RAN-side or CN-side information;
- **Network EE**: CN-Assisted RAN Energy Optimization Gain quantifies the improvement in RAN EE achieved by coordinating access traffic control in the CN, compared with a baseline where RAN energy optimization is performed without CN-side traffic coordination.

2 State of the Art

On the 3GPP side, Network Data Analytics Function (NWDAF) has evolved as a core analytics function for collecting, analyzing, and exposing network analytics for 5G system automation, while Management Data Analytics (MDA) provides management-plane analytics across performance measurements, KPIs, alarms, configuration data, Quality of Experience (QoE) reports, network analytics data, and service experience data [3GPP-NWDAF, 3GPP-MDA].

Recent implementations further integrate NWDAF with open-source 5GC functions, enabling UPF event exposure, ML lifecycle management, and SMF-driven closed-loop actions based on RT analytics [Shafiei2025]. On the O-RAN side, cross-domain AI is being framed through broader enablers such as data availability and quality, execution platforms, action mechanisms, inter-AI coordination, and governance, covering distributed/federated learning, model cooperation, trust management, and fallback operation [O-RAN-XAI].

In parallel, the O-RAN Next Generation Research Group (nGRG) has investigated service-based RAN and RAN-CN converged architectures as structural approaches to expose RAN capabilities, reduce control overhead, support AI/sensing capability exposure, enable edge convergence, and improve resource efficiency for 6G networks [O-RAN-SBRAN, O-RAN-RANCN]. Recent studies on service-based RAN also examine service definition, interface protocol stacks, radio capability exposure, and joint RAN-CN optimization, further supporting modular cross-domain coordination [SBA-RAN2025]. Looking toward 6G, LLM/agentic AI, DataOps/MLOps, and network DTs are emerging as key enablers for AI-native, trustworthy, and safe cross-domain automation, complementing NWDAF, RIC, and SMO-based control loops [3GPP-AI-NGRAN, AI-O-RAN2026].

3 New perspectives

DT for RAN technologies development

RAN technology development based on statistical channel models and channel abstraction has limitations in analyzing diverse real-world, particularly in generating large-scale datasets for AI/ML-based RAN development. 3D digital map-based simulations enabling diverse channel dataset generation, supporting effective AI/ML model training and intuitive performance evaluation.

AI Accelerators for RAN

RAN systems employing AI/ML technologies on AI accelerators (GPU/NPU) can enhance wireless transmission performance. Although AI accelerators still consume significant energy for training and operation, AI-based inference in RAN has the potential to improve long-term EE by achieving higher performance with the same radio resources or maintaining the same performance with reduced wireless resource usage.

Cross-domain optimization

As a new perspective, RAN/CN cross-domain coordination can integrate heterogeneous data and AI capabilities across traditionally separated domains to jointly optimize network operation and user experience. By combining RAN-side information such as channel conditions, mobility, interference, traffic load, and energy status with CN-side information such as session context, QoS flows, slice status, and service requirements, the network can support holistic AI-assisted control.

This enables operators to improve EE and resource utilization through coordinated resource allocation, traffic steering, and policy control, while also enhancing user service experience in terms of latency, reliability, mobility continuity, and application-level QoS.

4 Ethical Risks

From an ethics-risk perspective, RAN/CN cross-domain AI requires careful governance because it integrates sensitive user, service, mobility, and network data across multiple domains. Key concerns

include privacy leakage, excessive profiling, biased resource allocation among users or services, lack of transparency in AI-driven decisions, and accountability gaps when automated policies affect service quality.

To mitigate these risks, future systems should ensure data minimization, privacy-preserving analytics, explainable AI, fairness-aware optimization, secure data sharing, auditability, and reliable fallback mechanisms when AI decisions are uncertain or harmful.

3.3 AI-native 6G design for autonomous networking

This use case family embeds AI capabilities directly into the architectural foundation of 6G networks, transforming them from passive infrastructure into intelligent, self-governing systems. Unlike current approaches where AI serves as an external management tool, these networks will feature intelligence as a core design principle, enabling unprecedented autonomy and adaptivity across all network domains.

This represents a decisive shift from static, rule-based management to cognitive systems that continuously learn, anticipate demands, and proactively optimize resources. The use cases in this family demonstrate how this architectural revolution will create self-managing, energy-efficient networks capable of meeting the demands that will characterize the 6G era.

3.3.1 Anomaly Detection and Recovery Strategies

1 Summary

5GC networks are designed to deliver higher quality service in terms of speed and flexibility. But as service complexity and data traffic increase, it is ever more difficult to manage the services using fixed rules. This raises two main challenges. First, resources are inefficiently applied, which leads to wasted energy. Second, the service is more likely to be interrupted, which has a negative impact on the user experience. For instance, when traffic patterns change abruptly, the networks cannot react in a timely manner, causing either service interruptions or wasted resources. Also, when malfunctions occur, problem finding and fixing can take time leading to service delayed and extended periods of downtime.

Solving these challenges is critical to improving efficiency and consistency with service provision in 5G and future 6G networks. It is also a great way to positively impact business outcomes. More flexible and smarter resource management reduces energy usage with cost savings. A reduced service interruption improves the QoE. Faster responses to changing demands in traffic allow a company to remain competitive and deliver new services quicker. Improving network management thus becomes an important engine of growth and success.

1.1 Technical Description & Business Case

5GCs are designed to deliver high performance services with increasing flexibility and reliability, but the growth in traffic volume and service complexity makes static rule-based management progressively ineffective. In cloud native architectures defined by 3GPP TS 23.501 [23.501], the introduction of virtualization and microservice decomposition increases operational flexibility but also significantly raises system complexity. As a result, traditional management approaches struggle to react to dynamic conditions, leading to suboptimal resource allocation, higher energy consumption, and slower recovery from faults, which directly impacts service continuity and operational efficiency.

To address these limitations, modern 5GC management increasingly relies on AI driven techniques that exploit real time telemetry such as traffic patterns, signaling data, and QoS metrics. Predictive models like LSTM and GRU are used to forecast network demand, while RL optimizes resource allocation by scaling NFs and adjusting slice configurations. This enables a closed loop control

system where forecasting informs orchestration decisions, allowing the network to proactively reduce energy usage during low demand periods and rapidly restore capacity when traffic increases. In parallel, anomaly detection models such as autoencoders [NAVI_24] and Isolation Forest [GHANI_25] enhance resilience by identifying abnormal behaviour and enabling targeted corrective actions rather than full system resets.

Beyond core optimization, AI is also extended to multi operator environments and roaming scenarios, where RL supports dynamic network selection based on latency, load, and cost, while federated learning enables collaborative model training without sharing sensitive data. This combination of predictive control, anomaly detection, and distributed learning improves EE, operational resilience, and service quality, forming a clear business case for AI native 5G Core evolution aligned with sustainability and reduced operational expenditure.

1.2 KPIs and Requirements

1.2.1 KPIs and KVIs

The objective of this use case is to demonstrate how AI technologies can improve the efficiency, reliability, and automation of 5G network management. Predictive modelling for traffic forecasting and Natural Language Processing (NLP) for operational intelligence is leveraged to enhance the accurate interpretation of network conditions and support data-driven decision making. The KPIs defined below represent the target values that the proposed solution aims to achieve, providing a clear indication of its expected performance and effectiveness.

Table 6: Anomaly Detection and Recovery Strategies KPIs

KPI/KVI	Target or reported value
ARIMA MAE	< 0.05
ARIMA RMSE	< 0.1
NLP F1-Score	> 0.9
NLP Precision	> 0.9
NLP Recall	> 0.9

1.2.2 Functional and Non-Functional Requirements

No additional functional or non-functional requirements have been identified compared to those defined in the previous deliverable (D2.1), as the current solution continues to align with the established specifications.

2 State of the Art

5G networks have generated new security and dependability problems due to the use of virtualized infrastructure, many connected IoT devices, and complex, distributed architectures. Many recent papers have focused on using ML to detect unusual network behaviour and improve system adaptability. One method under investigation for anomaly detection without the need of decryption is CNNs trained on network traffic data. These approaches leverage AutoML tools and Neural Architecture Search (NAS) to maximize the model structure and convert traffic flows into inputs akin to images [JR20]. Another line of study proposes identifying 5G RAN problems including signal interference or power reduction using Transformers in concert with Graph Neural Networks (GNNs) [ACK24]. This approach models the temporal and spatial interactions between base stations thus improving failure detection and root cause analysis. Finally, as a 5G CN behavior changes with time,

that can cause model drift. To address that issue several approaches have been proposed, for instance [DAA22] proposes a framework that automatically retrains the predictive model in response to traffic pattern changes, thus helping maintain or restore the model's accuracy. These studies highlight significant problems including the lack of real, labelled data, the likelihood of insufficient generalization to fit real-world situations, and the need for balancing system efficiency with detection accuracy.

3 New perspectives

No new perspectives have been identified with respect to the previous deliverable (D2.1).

4 Ethical Risks

Firstly, the implementation of AI-driven network management for 5GC poses risks in terms of data processing, automated decisions, and their accuracy. Because traffic prediction, anomaly detection, and roaming optimization depend on the continuous analysis of network telemetry, there is a possibility that this process might result in indirect disclosure of users' or services' sensitive data. Such risk can be avoided by implementing appropriate data minimization policies, anonymization, and privacy-preserving technologies, such as federated learning, which ensures that no data is directly exchanged between service providers.

Secondly, automated decision-making may be used to allocate resources and restore connections, but faulty predictions or models may cause ineffective energy usage, discriminatory priorities for services or users, and forced disruption of services. To avoid possible negative effects, AI decisions need to be easily audited and explained to be subjected to policy restrictions.

Finally, automation increases the risks associated with misconfigured management systems, potentially resulting in large-scale disruption of services. In order to minimize such risks, proper operational limits should be imposed, model validation needs to be performed regularly, and a fallback mechanism in case of system failure should be put in place.

3.3.2 Cultivating Radio Measurements into Value-Added Services for Spectrum Sharing

1 Summary

In this use case the data that network operators collect to ensure the performance of the network and proper delivery of the service is enhanced and processed to improve resource usage and to enable entirely new services. The value of this refined data is not limited only to the operator that collected and processed it; rather, it is an asset that can be exchanged and traded between different actors of the network value chain as well as other operators sharing the same service locations. Such information can be aggregated and captured into "service usage maps" that reflect details like QoS demand for specific times and locations and can be shared with other parties such as Smart City DTs for further analysis and actions.

1.1 Technical Description & Business Case

Consider a system composed of Wi-Fi and New Radio Unlicensed (NR-U) access technologies, which share a frequency band in a specific location or area and could therefore exchange forecasted service usage maps to manage shared spectrum resources in a collaborative manner. This approach is essential for businesses and industries deploying their own private networks (e.g., for smart factories, ports, or other localized applications) without the need to acquire exclusive license [MYS+19] or deploying costly sensors.

In 6G, the NWDAF is evolving into a Distributed Intelligence Function that actively controls the network. It will not just predict congestion but autonomously reroute traffic and reconfigure radio resources to meet the SLAs. The insights and predictions provided by the 6G NWDAF can be further processed into value-added information for external services and service providers (e.g., DTs, anomaly detection, spectrum sharing, etc., see more about potential applications and the value proposition in D2.1). The exact information contained in service maps depends on the services considered (e.g., URLLC, precision sensing, ubiquitous computing), such as highly accurate location information, usage patterns, mobility pattern and beam management parameters, etc.

For the specific case of Wi-Fi and NR-U coexistence, radio measurements such as CSI and sounding data are collected and enhanced with service-level information such as QoS levels and Joint Communication and Sensing (JCAS) data. This data is used to generate specific radio maps and channel charts, which are then exchanged between collaborating operators that coexist and share resources on the same frequency bands, such as the sub-7GHz spectrum. Sharing radio maps is expected to facilitate more dynamic and efficient radio resources and traffic management in overlapping coverage areas. It is possible to apply this very same approach of sharing radio maps between AI agents or controllers of different cellular radio technologies (i.e. between 5G and 6G) and thus enhance Multi-Radio Spectrum Sharing (MRSS) operations with the use of predictive radio maps.

1.2 KPIs and Requirements

1.2.1 KPIs and KVIs

Relevant KPIs for this use case are included in section 2.2.2. The main KPI is “spectrum and bandwidth” followed by “area traffic capacity”, “peak data rates”, “U-plane latency”, “network and terminal EE”, and “mobility”.

1.2.2 Functional and Non-Functional Requirements

Functional requirements:

- Access to Wi-Fi and 6G/NR-U CSI information must be provided;
- An exposure API between Wi-Fi (management or agent) systems and NR-U radio resource controller (or AI-agent) is required to exchange spectrum usage maps. This could take place via a cross RAT signaling interface between Wi-Fi APs and 6G base stations, or, alternatively, federated learning can be applied to ensure stronger privacy.

Non-Functional requirements:

- Security and Privacy: The information exchanged between actors (i.e., operators) across the domains must comply with data protection regulations, such as the GDPR. The system must maintain high trust, compliance, and risk mitigation through consistent enforcement of security policies and encryption standards. End-users privacy must be prioritized;
- Accuracy: The exchanged information must be accurate and timely considering the applications and services used;
- Tailored service maps should reduce unnecessary training, deployment and maintenance of the AI-agents involved and their AI/ML models by leveraging preprocessed data contained in the service maps;
- Tailored service maps should improve the end-user experience.

2 State of the Art

In 5G, NWDAF [23.288] plays a central role in optimizing network performance and improving network efficiency. It collects data from UE, NFs, and OAM systems. 3GPP Releases 16 and 17 expanded its capabilities to include support for federated learning and data management.

The 6G NWDAF (or evolved NWDAF) will evolve from a centralized to a distributed intelligence architecture where intelligence is embedded in every NF and potentially even in the UEs, allowing for local real-time decision-making [SGO+24, LS25]. 6G NWDAF will support federated learning that allows models to be trained across multiple domains or devices without sharing raw data, thereby enabling cross-domain insights while maintaining strict privacy [YHZ25]. 6G NWDAF is also expected to feed real-time network data into DTs.

The use cases in [22.870] do not cover spectrum sharing directly, but there are a few use cases that could benefit from the presented scenario. Namely, “the use case on 6G Agent collaboration with third party using LLM”, “the use case on AI/ML model training and inference” and “the use case on 6G native AI in multi-domain coverage”.

The "Study on 6G Scenarios and Requirements”, TR 38.914, explicitly covers 6G MRSS scenarios and requirements for spectrum sharing with 5G NR but it doesn’t mention the use of service specific radio maps exposed between cooperation networks.

3 New perspectives

Using distributed federated NWDAF between 5G and 6G with the capability of exchanging predictive service specific radio maps would complement the MRSS approach.

4 Ethical Risks

The main ethical risk is related to monitoring physical environments and radio signal dynamics that may reveal human behaviour. 6G sensing data (CSI, sounding data) is so granular that even if the UE Identifiers (IDs) are removed, "mobility patterns" and "indoor localization" can create a unique behavioural fingerprint. Traditional data collection requires UEs to be connected and a consent can be requested, however, if JCAS are to be used, it would be technically possible to detect a person’s presence or movements inside a private building (indoor tracking) using the reflection of 6G signals off their body without their consent.

A second identified ethical risk is related to the possible grouping of UEs (and users) based on what services they are using. This might inadvertently create a "two-tier" internet based on socio-economic indicators derived from collected performance and location data leading to potential discrimination if not properly regulated.

4 Mapping the Use Cases to the 6GARROW Initial Architecture

To fulfil the requirements of the 6GARROW use cases summarized in Table 1, the 6GARROW system architecture was designed to provide support for semantic-aware edge learning and inference, integrating RAN & CN aspects. In principle, the architectural design should incorporate several functional requirements that stem from the key challenges associated with edge learning and inference, such as resource constraints, scalability and heterogeneity, orchestration of intelligence, and decentralized learning. Under the proposed framework, the device (e.g., a UE) and the network would collaborate to perform edge learning and inference tasks according to criterions that capture trade-offs between KPIs associated with AI/ML capabilities, such as training complexity, inference accuracy, AI/ML-related communication overhead, model generalization capability, AI/ML performance, and inference latency.

In the following, we revisit the initial 6GARROW functional architecture presented in D2.2 and provide an initial mapping between the use cases and the architecture, based on the functional requirements associated with each of the use cases.

4.1 6GARROW Initial Architecture

The initial, use case driven, functional 6GARROW architecture was presented in D2.2. The architecture emphasizes the AI-native capabilities supporting joint RAN–CN optimizations with the aim of efficiency and sustainability. The functional architecture was derived based on the use cases described in D2.1 and it groups different AI-native functionalities in different subsystems and provided initial information flows between the functional groupings.

In what follows, we revisit that initial architecture to take on board some recent developments related to the work in the project, to some news from the advancement of the state of the art, and to the continuously evolving standardization work landscape.

The main changes in comparison to the earlier version in D2.2 are the addition of the AI-native Link and Flow control layer in the UE and in the RAN, the replacement of AI model and agent deployment on the top in the “AI stack” on the same level with MLOps, and the shift of the loss and parameter control into the AI-control layer, as shown in the following figure.

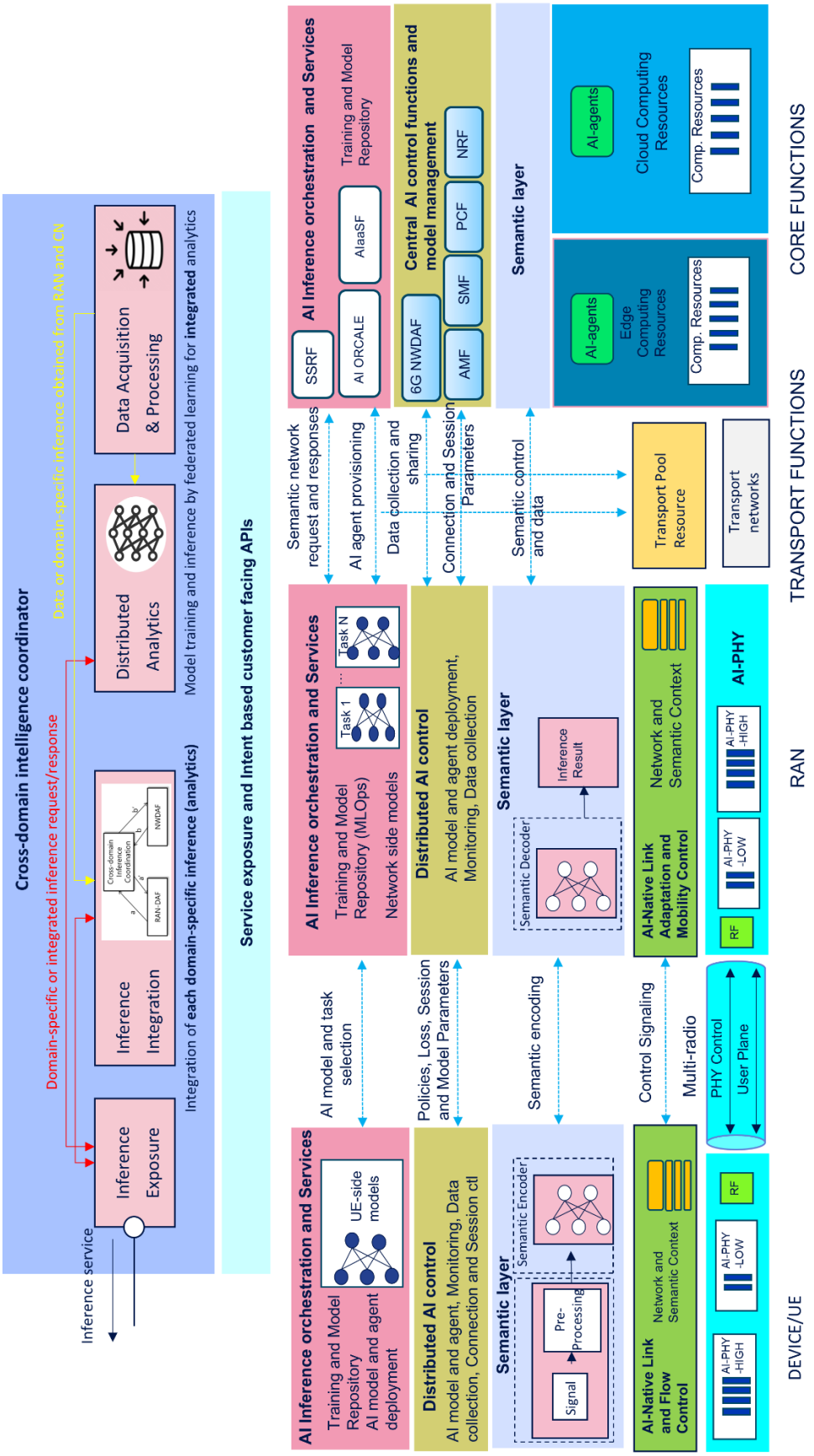

Figure 7: Revisited functional 6GARROW architecture.

The new version of the 6GARROW functional architecture can be summarized as a layered architecture as show in Fig 8.

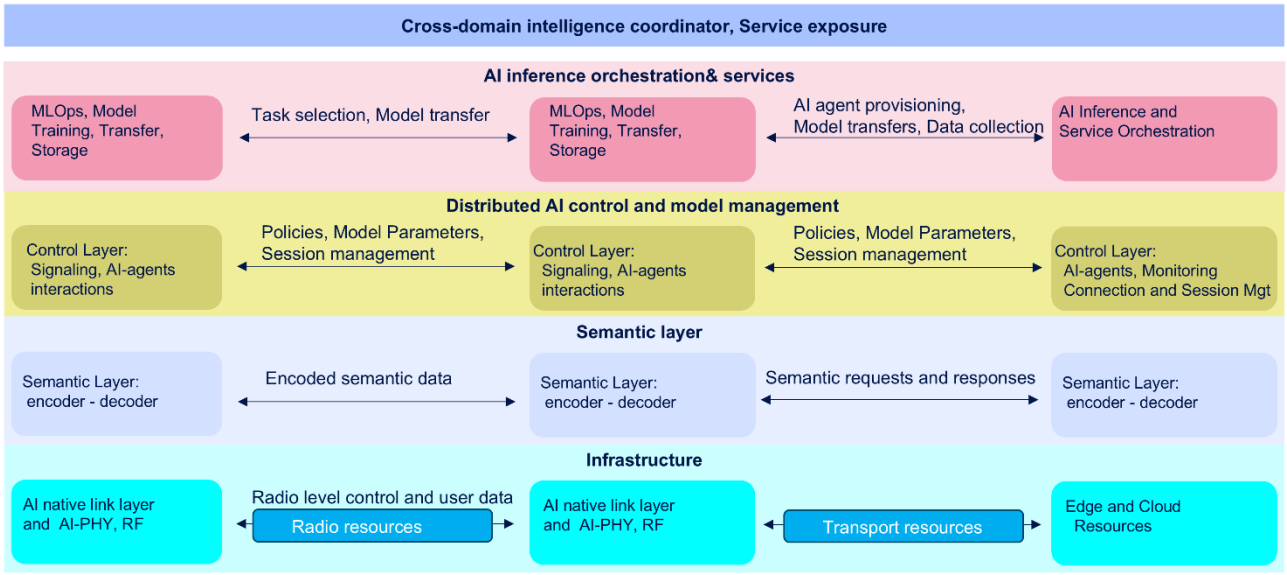

Figure 8: 6GARROW High-level view on the functional architecture.

The infrastructure layer, encompassing radio, transport and cloud resources, is expected to be AI-enabled to provide full flexibility in resource and task orchestration for joint RAN-CN optimizations.

The semantic layer contains three enablers: (i) non-arbitrary semantic representations, (ii) dedicated interfaces to negotiate semantic-centric SLAs or SSLAs and exchange associated information and metrics, and (iii) a semantic control plane to configure network functions accordingly. Moving towards semantic communication networks involves adopting a new, data representation that can be managed by ML models to meet diverse application-specific requirements.

The Distributed AI Control and Management layer covers AI model and AI agent runtime operations, such as model selection, policy control, and data collection.

The AI Inference orchestrations and services layer manages the deployment and execution of models for prediction and exhibits service exposure towards external actors.

The Service layer exhibits service exposure and intent-based services towards external actors and cross-domain coordination.

4.2 Mapping of 6GARROW use cases to the functional architecture

Each use case family targets a specific dimension of the AI-native and device-integrated 6G architecture, demonstrating how these concepts can be operationalized to deliver tangible benefits across a variety of application domains, as show in Table 7.

Table 7: Mapping of the use cases to the 6GARROW functional architecture

	Use case/ Essential architectural requirements	Cross-domain coordination	AI Inference and Services	Distributed AI control and model management	Semantic layer	Infrastructure
AI native 6G design empowering new network services for stakeholders	AI-as-a-Service for stakeholders & creation of tailored AI models according to End Users needs	Must have an interface connecting intra-network entities to extra-Network (for example Cloud) entities for trusted processing of MNO-internal data on external (Cloud) entities.	Must have an interface enabling access to AI/Data Management entity to request creation and/or refinement of AI models	Must have an interface enabling access to AI/Data Management entity to request creation and/or refinement of AI models		
	URLLC for Machine Vision-based Robotic Control	A robust remote-control Application Programming Interface (API), providing cloud applications with an abstract interface to issue high-level semantic commands	Coordination of an AI-driven orchestration and scheduling framework for semantic interactions and extraction of intents		Must integrate URLLC iwith semantic intelligence.	A semantic communication protocol stack is required to extract only the essential meaning of multimodal data
	Semantic Communications Services for Agents			Security and Access Control for AI-agents	Adapted protocol stacks specifically designed for embedding-based transmission patterns	Should enable efficient workload distribution across computing nodes
Resource Usage (lower layers)	Highly efficient communications with light infrastructure				The system should be capable of extracting the essential meaning or context of the transmitted data at both the sender and receiver ends.	Semantic Encoding: Only the meaningful components that contribute to the understanding of the message are transmitted. Context-Aware Quality of Service
	Dynamic resource allocation for intercontinental RAN-Core	A Cross-Domain AI Orchestration Framework is needed, led by CNN, VAE, and Transformer models, to dynamically orchestrate bandwidth, latency, and processing resources across international links.		Must support forecast based resource scheduling	Autoencoders must monitor deviations in traffic or latency patterns, allowing early detection and correction of failures or inefficiencies.	
	Overhead reduction through Semantic Closed Control Loop for Private 5G Management Systems			Must collect and combine various types of data from different sources within the mobile network. Data is linked clearly with network intents and that actions are triggered correctly through an orchestration system.		
	RAN/UE & RAN/CN Cross-Domain Coordination		Must implemnt AI/ML-based Inference for CSI prediction and beam management			Must conform to interface specifications defined by the 3GPP RAN specification to ensure interoperability and seamless integration within standardized RAN architectures.
Automation & Failure Recovery (network functions)	Anomaly Detection & Recovery Strategies			Must parse network alerts through NLP to comprehend the context of alerts,		Must monitor UPF throughput in real time continuously, make predictions regarding traffic patterns, and scale UPF instances
	Cultivating Radio Measurements into Value-Added Services for Spectrum Sharing	Must provide access CSI info across domains. Must have an exposure APIs for cross domain radio map exchange		Must provide access to CSI or similar data		

4.3 Connection to the WP5 Activities

6GARROW will demonstrate key architectural innovations and will validate novel AI/ML solutions for enhanced device performance and optimized RAN and CN. Going beyond the functional demonstration and validation, some demonstrations will be developed to showcase an AI-enhanced network where AI functionalities are integrated across all domains (device, RAN and CN). In addition, the Proof of Concept (PoC) will demonstrate advanced orchestration of network functions, such as fronthaul/backhaul management, authentication, and user management, while enabling an open interface for customer-driven application development, implementation, integration and testing of testbeds for functional demonstrations.

In the following we summarize the planned demonstrations and the PoC, and we provide relations between the demonstrated concepts and the blocks of the enhanced functional architecture discussed in Section 4.1.

4.3.1 Planned Demonstrations and PoC

Demonstration 1: Semantic-aware device-edge co-inference for human-robot interaction. This demonstration will deliver an implementation of an architecture for device-edge co-inference that integrates semantic-aware encoding and inference with wireless communication. The demonstration will be based on the testbed and experimentation platform for edge intelligence applications, currently under development at FhG. A concrete example is provided by the demo setup illustrated in Figure 9, which implements gesture recognition in a robotic control application. The depicted setup integrates sensing, on-device semantic encoding performing, NR-based transmission/reception over a wireless channel, and semantic decoder for edge inference.

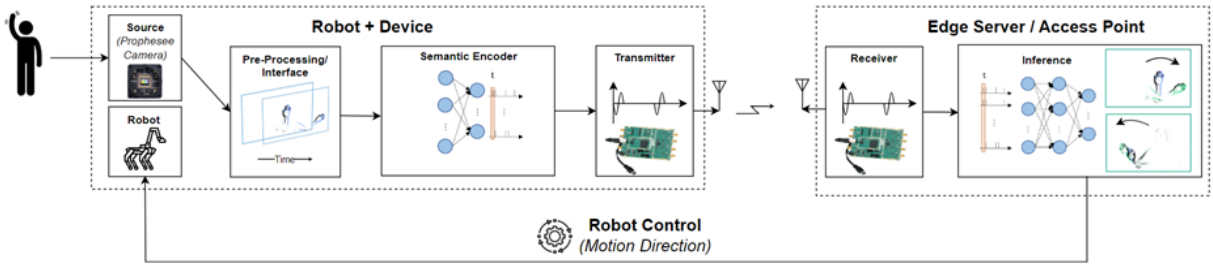

Figure 9: Semantic aware device-edge co-inference for robotic control.

Demonstration 2: 6G cross-domain network intelligence framework
The focus of this demonstrator is to showcase inference coordination (see Fig. 10) for cross-domain network slicing for E2E AI-native 6G networks. Components of the demonstration include: (i) network analytics functions from the CN domain performing data collection, training and inference to deliver network statistics and prediction; (ii) network analytics functions from the RAN domain to efficiently manage and optimize network resources, deliver differentiated services, and meet the diverse requirements of the use cases; (iii) exposed analytics related to network slices from RAN and CNs; and (iv) cross-domain inference coordination function to validate whether the requirements can be met under the current condition.

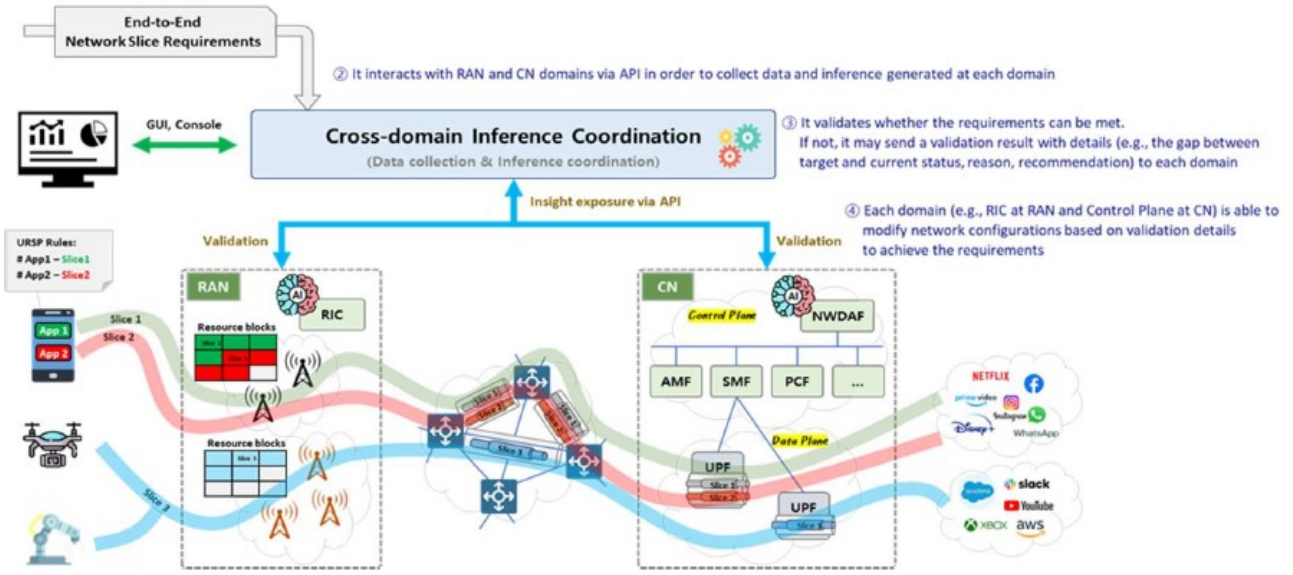

Figure 10: Illustration of the demonstration setup "Inference coordination for cross-domain network slicing".

Demonstration 3: Physical layer AI/ML techniques

This demonstration (Fig. 11) is designed to assess the effectiveness of AI/ML PHY techniques such as CSI feedback and beam management. The demo system will be constructed using O-RAN-defined components, namely O-RU, O-DU, and O-CU.

This lab demo shows potential in validating the performance and efficiency of AI/ML algorithms specifically tailored for the PHY within realistic network environments.

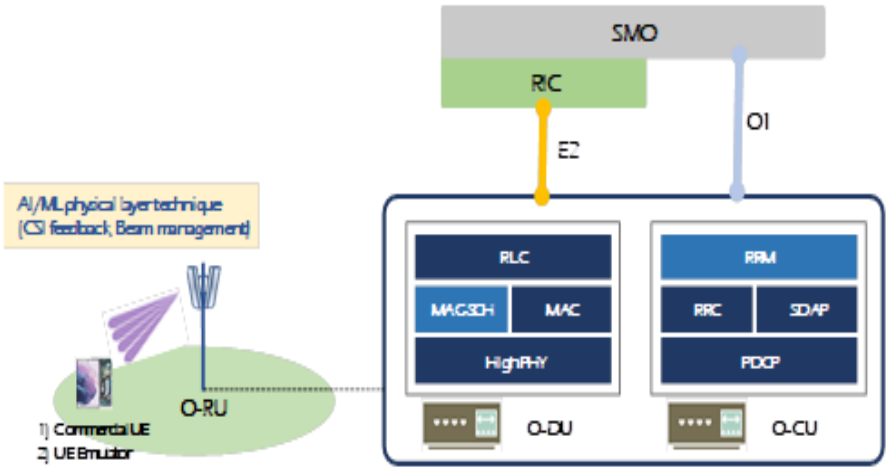

Figure 11: Demonstration of physical layer AI/ML techniques.

Demonstration 4: AI/ML based CSI and CQI compression

This demonstration will provide an implementation of the compression of CSI and CQI algorithms in an edge intelligence platform. The platform is based on an existing testbed including Xilinx RFSoC architecture. The testbed will be enhanced by adding edge intelligence with dedicated GPUs, as illustrated in Fig. 12. The validation of the algorithms performance will be done in two steps: (i) with a channel emulator; and (ii) with from field channel measurement.

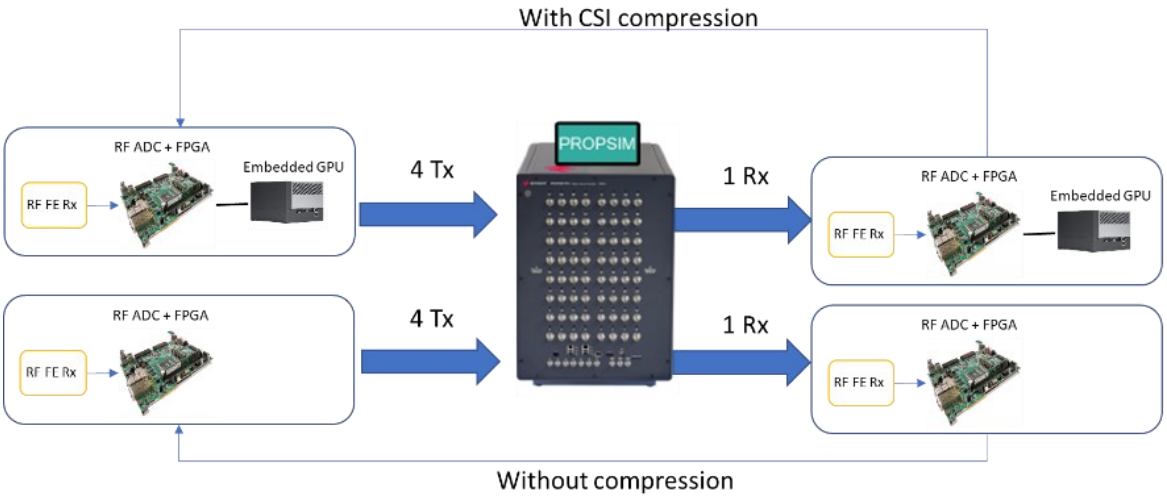

Figure 12: Platform for CSI/CQI compression assessment.

Intercontinental PoC: Development of a joint testbed system and experimental evaluation of the key AI/ML concepts

The joint PoC demonstrates AI-enabled procedures to improve wireless efficiency through mobility/resource management, maintenance, and self-optimization of network parameters. It follows an evolved architecture merging RAN and CN into a unified system managed by radio and network controllers. The PoC utilizes 5G test networks at Yonsei University in South Korea and Aalto University in Finland.

Two CNs will be used: HPE's 5GC and Cumucore's CN available at Aalto and Yonsei. This enables testing of mobility management in large scale including roaming. The PoC system will use virtualized RAN at Yonsei campus and more traditional RAN architecture with RIC at Aalto campus. An AI-powered Network Controller orchestrates CU functions, manages fronthaul and backhaul, handles authentication, network slicing, and user management, and provides open interfaces for third-party application development.

The PoC will also integrate aspects of the semantic-aware device-edge co-inference architecture, showcasing the potential of 6G to provide overhead-aware, low-latency, energy-efficient solutions for edge intelligence applications. HPE contributes a fully virtualized, 3GPP and ETSI-compliant 5G CN including key functions: AMF, SMF, UPF, UDM, AUSF, UDR, PCF, NRF, NSSF, and EIR. It supports standard interfaces (N1–N6) and allows trusted third-party applications to configure network parameters and control E2E QoS for mission-critical services.

4.3.2 Mapping between the Demonstrations and the Functional Architecture

In Table 8 we provide a mapping between the planned demonstrations and the 6GARROW functional architecture discussed in Section 4.1. In addition, we highlight the use case families addressed by each of the planned demonstrations.

Table 8: Mapping between the demonstrations, the use cases and the functional architecture

Demonstrations	Use Case Families			Cross-domain coordination and service exposure	AI inference orchestration and services	Distributed AI control and model management	Semantic layer	Infrastructure
	F1	F2	F3					
Demonstration 1	✓					✓	✓	✓
Demonstration 2		✓		✓				✓
Demonstration 3			✓					✓
Demonstration 4		✓			✓	✓		✓
Intercontinental Demonstration	✓	✓	✓	✓	✓	✓	✓	✓

* Use Case Family 1 (F1): AI native 6G design empowering new network services for stakeholders.

* Use Case Family 2 (F2): Resource usage (lower layers).

* Use Case Family 3 (F3): Automation & failure recovery (network functions).

* In the context of the Use Case Families, "✓" indicates that the demonstration addresses aspects present in some of the Use Cases belonging to the Use Case Family.

* In the context of the Functional Architecture, "✓" indicates that the demonstration involves concepts in the domain of the corresponding block/layer of the Functional Architecture (see Fig. 8).

5 Conclusion

This deliverable provides an improved view of the 6GARROW scenarios, use case families, and their respective KPIs and KVIs that contribute to the AI-native 6G networks vision. As a continuation of D2.1, D2.3 further clarifies the understanding of use cases and performance indicators while incorporating recent advances in state-of-the-art research and standardization efforts, ethical risks and new perspectives, where applicable.

In addition, D2.3 revisits the functional architecture presented in D2.2 and provides an initial mapping of the use cases to the 6GARROW functional architecture, based on the functional requirements associated with each use cases.

One of the primary achievements of this deliverable is the more explicit connection between the improved use case requirements, KPIs and KVIs, and architectural enablers. Through the detailed identification of relationships between use cases and the 6GARROW functional architecture, D2.3 allows transitioning from general use-case scenarios to more specific functional components. In addition to evaluating technical performance, the improved KPI/KVI framework covers operational efficiency, power consumption, and business aspects.

The refined use cases, scenarios, and KPIs/KVIs described in this deliverable will serve as inputs for the technical efforts within WP3, WP4, and WP5, namely development and validation of 6GARROW's PoC.

D2.3 will be instrumental in the design and evaluation of new AI-based mechanisms and platforms for achieving the envisioned future of the telecommunication system, as proposed by 6GARROW.

References

[22.261] 3GPP, Service Requirements for the 5G System, document TS 22.261 v16.0.0, 3rd Generation Partnership Project, Jun. 2017.

[22.870] 3GPP TR 22.870 “Study on 6G Use Cases and Service Requirements”, Technical Report, V20.0.0, March 2026.

[221348] 3GPP RP-221348, Revised SID: Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR Air Interface, June 2022.

[23.288] 3GPP TS 23.288 V20.0.0, Architecture enhancements for 5G System (5GS) to support network data analytics services (Release 16), March 2026.

[23.501] 3GPP TS 23.501 V17.0.0, “System Architecture for the 5G System (Release 17),” Dec. 2023.

[23.530] 3GPP SA5, TS 23.530 Management and orchestration; Concepts, use cases and requirements v19.0.0 (2025-03).

[23.533] 3GPP SA5, TS 23.533 Management and orchestration; Architecture framework v19.1.0 (2025-03).

[28.801] 3GPP TR 28.801 V16.0.0, “Study on energy efficiency metrics for the E-UTRA and E-UTRAN,” Jun. 2022.

[29.520] 3GPP, “5G; 5G System; Network Data Analytics Services”, Accessed on 03/04/2025[Online]. Available: https://www.etsi.org/deliver/etsi_ts/129500_129599/129520/15.03.00_60/ts_129520v150300p.pdf.

[36.888] 3GPP TR 36.888, “Study on provision of low-cost Machine-Type Communications (MTC) User Equipments (UEs) based on LTE (Release 12)”, June 2013.

[38.913] 3GPP TR 38.913 V17.0.0, “Study on scenarios and requirements for next generation access technologies,” Mar. 2021.

[3GPP+23] 3GPP SA5, ‘Intent Driven Management’, 3GPP technology overview, 2023.

[3GPP-AI-NGRAN] 3GPP, “AI/ML for NG-RAN & 5G-Advanced towards 6G,” 3GPP Technology Trends, 2026.

[3GPP-NWDAF] 3GPP, “Network Automation Enablers in 5GS,” 3GPP Technology Trends, 2025.

[3GPP-MDA] 3GPP TS 28.104, “Management and orchestration; Management Data Analytics (MDA),” Release 17/18.

[6GARROW-D21] 6GARROW, Deliverable D2.1 “Scenarios, Use Cases and related KPIs/KVIs”, 31/08/2025.

[6GARROW-D22] 6GARROW, Deliverable D2.2 “Initial System Architecture”, 31/10/2025.

[6GARROW-D31] 6GARROW, Deliverable D3.1 “State of the art and challenges for AI/ML enhanced device performance”, 31/10/2025.

[6GARROW-D41] 6GARROW, Deliverable D4.1 “State of the art on AI/ML Solutions for RAN and Core Optimization” 31/10/2025.

[6GGOALS] Project homepage. [Online]. Available: https://sites.google.com/view/6ggoals/home.

[6GWS-250238] 6GWS-250238 AI# 6 Chair’s summary of the 3GPP Workshop on 6G, Mar., 2025.

[6GIA24] 6G Industry Association, European Vision for the 6G Network Ecosystem, 6G-IA White Paper, Nov. 2024. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2021/06/White-Paper-6G-Europe.pdf.

[A2A] “Unlock Collaborative Agent Scenarios”. [Online]. Available: https://google.github.io/A2A/.

[AAC24] Abdelkader Mekrache; Adlen Ksentini; Christos Verikoukis, “Intent-Based Management of Next-Generation Networks: an LLM-Centric Approach”, IEEE Network (Volume: 38, Issue: 5, September 2024).

[ABB+22] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar et. al., “Do as I can, not as I say: Grounding language in robotic affordances,” arXiv preprint arXiv:2204.01691, 2022.

[ACK24] Antor Hasan, Conrado Boeira, Khaleda Papry, et al., “Root Cause Analysis of Anomalies in 5G RAN Using Graph Neural Network and Transformer”, arXiv:2406.15638v1.

[AI_Act_Explorer_26]. “AI Act Explorer”. Available only at: https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-5.

[AI-O-RAN2026] “The Architecture of AI and Communication Integration towards 6G: An O-RAN Evolution,” 2026.

[B5G22] Beyond 5G Promotion Consortium, Message to the 2030s, White Paper, Mar. 2022. [Online]. Available: https://b5g.jp/doc/whitepaper_en_1-5.pdf.

[BGV+22] A. Bayen, J. Gao, A. Velu, E. Vinitsky, Y. Wang, Y. Wu, and C. Yu, “The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games,” in *Advances in Neural Information Processing Systems 35*, ser. NeurIPS 2022, no. PPO. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022, pp. 24 611–24 624.

[Blackridge_report_25]. “NVIDIA and Intel Announce $5 Billion Partnership to Develop AI Infrastructure and Personal Computing Products“. Available online at: https://www.blackridgeresearch.com/news-releases/nvidia-intel-announce-5-billion-partnership-develop-ai-infrastructure-pc-products-us.

[C26.0.933] TM Forum Catalyst C26.0.933, “Zero Trust Agent Reference Architecture — Telco”, v2.4. [Online]. Available: https://www.tmforum.org/catalysts/projects/C26.0.933.

[Chumien_25] W. Chumyen, "Multi-Agent Smart Water Distribution with Computer Vision-Based Container Detection and AI-Driven Behavior Forecasting for Disaster Relief in Smart Cities," *2025 IEEE International Conference on Agentic AI (ICA)*, Wuhan, China, 2025, pp. 7-11, doi: 10.1109/ICA67499.2025.00011.

[CPU_Intel_26] “AI Inference Acceleration on CPUs”. Available online at: https://www.intel.com/content/www/us/en/developer/articles/technical/ai-inference-acceleration-on-intel-cpus.html.

[DAA22] Dimitrios Michael Manias, Ali Chouman, Abdallah Shami, “A Model Drift Detection and Adaptation Framework for 5G Core Networks”, arXiv:2209.06852v1.

[DAA25] Dimitrios Michael Manias, Ali Chouman, and Abdallah Shami, "Semantic Routing for Enhanced Performance of LLM-Assisted Intent-Based 5G Core Network Management and Orchestration.", arXiv 2404.15869v1.

[Dataeconomy_report_26] “China targets 2030 for trial commercial 6G deployment”. Available online at: https://dataconomy.com/2026/03/30/china-targets-2030-for-trial-commercial-6g-deployment/.

[DIN23] Ding, Guangyao, et al. "Joint urllc traffic scheduling and resource allocation for semantic communication systems." IEEE Transactions on Wireless Communications 23.7 (2023): 7278-7290.

[GHANI_25] E. P. Ghani, A. Sofwan and M. Somantri, "AI-Driven Network Security: Detecting and Mitigating DDoS, Malware, and Backdoor Attacks with Isolation and Random Forest Algorithm," *2025 International Conference on Smart Computing, IoT and Machine Learning (SIML)*, Surakarta, Indonesia, 2025, pp. 1-6, doi: 10.1109/SIML65326.2025.11080951.

[GDW+23] Y. Gu, L. Dong, F. Wei, and M. Huang. “MiniLLM: Knowledge distillation of large language models.” arXiv preprint arXiv:2306.08543, 2023.

[GovI23] Government of India, 6G Vision, Department of Telecommunications, Bharat White Paper, Mar. 2023.[Online]. Available: https://dot.gov.in/sites/default/files/Bharat%206G%20Vision%20Statement%20-20full.pdf.

[HEX23-D52] Hexa-X-II Deliverable D5.2 Characteristics and classification of 6G device classes. 2023.A18.

[HEX23-D43] Deliverable D4.3 AI-driven communication & computation co-design: final solutions, Hexa-X Project, 2023, available at https://hexa-x.eu/wp-content/uploads/2023/05/Hexa-X_D4.3_v1.0.pdf.

[HPE] HPE, “HPE Aruba Networking Private 5G”, Accessed on 20/03/2025.[Online]. Available: https://www.hpe.com/it/it/aruba-networking-private-5g.html.

[HPZ+25] Q. Hou, S. Park, M. Zecchin, Y. Cai, G. Yu, and O. Simeone, “What If We Had Used a Different App? Reliable Counterfactual KPI Analysis in Wireless Systems,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 11, no. 5, pp. 3529–3543, Oct. 2025.

[IG1453] IG1453 Agent to Agent Protocol for Telecoms (A2A-T) v2.0.0.

[IG1453A] IG1453A Structured Prompt of Agent-to-Agent Protocol for Telecoms (A2A-T) v1.0.0.

[IMT23a] IMT-2030 (6G) Promotion Group, 6G Usage Scenarios and Key Capabilities, White Paper, June 2023.[Online]. Available: https://www.imt2030.org.cn/html/default/en/Publications/Whitepaper/index.html?index=2.

[IMT23b] IMT-2030 (6G) Promotion Group, 6G Wireless System Design Principles and Typical Features, IMT-2030 (6G) PG, Doc. 024,[Online]. Available: https://www.imt2030.org.cn/html/default/en/Publications/Whitepaper/index.html?index=2.

[ISP25] Joseph H R Isaac, Harish Saradagam, Nallamothu Pardhasaradhi, ‘5G core fault detection and root cause analysis using machine learning and generative AI’, arXiv preprint repository, 2025, arXiv 2508.09152.

[JR20] Jordan Lam, Robert Abbas, “Machine Learning based Anomaly Detection for 5G Networks”, arXiv:2003.03474v1.

[JS21] S. Javaid and N. Saeed, “Carbon-Aware Orchestration of Integrated Satellite Aerial Terrestrial Networks via Digital Twin.” arXiv:2510.17825, 2025.

[KBZ+24] V. Kasuluru, L. Blanco, E. Zeydan, A. Bel, and A. Antonopoulos, “Enhancing Cloud-Native Resource Allocation with Probabilistic Forecasting Techniques in O-RAN,” in *2024 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*. IEEE, 2024, pp. 1–6.

[Kim_24] J. Kim, J. Jeon, J. Park and S. Jung, "Continuous Performance Improvement of Infrastructure Guidance Service for Autonomous Cooperative Driving: Focusing on Data-centric AI," *2024 International Conference on Electronics, Information, and Communication (ICEIC)*, Taipei, Taiwan, 2024, pp. 1-4, doi: 10.1109/ICEIC61013.2024.10457156.

[KS21] N. Kazemifard and V. Shah-Mansouri, “Minimum Delay Function Placement and Resource Allocation for Open RAN (O-RAN) 5g Networks,” *Computer Networks*, vol. 188, no. technical, p. 107809, Apr. 2021.

[KTD+25] A. Koochali, E. Tahaei, A. Dengel, and S. Ahmed, “VAEneu: A Nnew Avenue for VAE Application on Probabilistic Forecasting,” *Applied Intelligence*, vol. 55, no. 7, Feb. 2025.

[LAL+21] B. Lim, S. O. Arık, N. Loeff, and T. Pfister, “Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting,” *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, Oct. 2021.

[LDS+24] C. Liang, H. Du, Y. Sun, D. Niyato, J. Kang, D. Zhao, and M. A. Imran. “Generative AI-driven semantic communication networks: Architecture, technologies and applications.” IEEE Transactions on Cognitive Communications and Networking, early access, 2024.

[LLZ+21] K. Lin, Y. Li, Q. Zhang, and G. Fortino, “AI-Driven Collaborative Resource Allocation for Task Execution in 6G-Enabled Massive IoT,” *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5264–5273, Apr. 2021.

[LS25] M. Lima Romero M., R. Suyama R., Towards Network Data Analytics in 5G Systems and Beyond, XLIII Simpósio Brasileiro de Telecomunicações e Processamento.

[LSX+25] Lam Dinh, Sihem Cherrared, Xiaofeng Huang, and Fabrice Guillemin, "Towards End-to-End Network Intent Management with Large Language Models", arXiv 2504.13589v1.

[LSZ+24] S. Li, Y. Sun, J. Zhang, K. Cai, H. Chen, S. Cui, and X. Xu. “Cooperative semantic knowledge base update policy for multiple semantic communication pairs”, arXiv preprint arXiv:2410.02405, 2024.

[M.2410-0] Report ITU-R M.2410-0 (11/2017): Minimum requirements related to technical performance for IMT-2020 radio interface(s).

[MDC+24] M. Martinez-Morfa, C. R. De Mendoza, C. Cervello-Pastor, and S. Sallent, “DRL-based xApps for Dynamic RAN and MEC Resource Allocation and Slicing in O-RAN,” in *2024 15th International Conference on Network of the Future (NoF)*. IEEE, Oct. 2024, pp. 106–114.

[MFW23] X. Ma, G. Fang, and X. Wang. “LLM-pruner: On the structural pruning of large language models.” In Proc. NeurIPS, New Orleans, USA, Dec. 2023, pp. 21702–21720.

[MKP23] Kashif Mehmood, Katina Kralevska, David Palma, ‘Intent driven autonomous network and service management in future cellular networks A structured literature review’, Computer Networks, Elsevier, 2023, DOI 10.1016/J.COMNET.2022.109477.

[MPS21] J. P. Mattsson, B. Smeets, and E. Thormarker. “Quantum-resistant cryptography.” arXiv preprint arXiv:2112.00399, 2021.

[MPY+25] Md Arafat Habib, Pedro Enrique Iturria Rivera, Yigit Ozcan, Medhat Elsayed, et al., "Harnessing the Power of LLMs, Informers and Decision Transformers for Intent-driven RAN Management in 6G", arXiv 2505.01841v1.

[MYS+19] M. Matinmikko-Blue, S. Yrjölä, V. Seppänen, P. Ahokangas, H. Hämmäinen and M. Latva-Aho, "Analysis of Spectrum Valuation Elements for Local 5G Networks: Case Study of 3.5-GHz Band," in IEEE Transactions on Cognitive Communications and Networking, vol. 5, no. 3, pp. 741-753, Sept. 2019, doi: 10.1109/TCCN.2019.2916309.

[NAVI_24] S. Navyakala and R. Rathinam, "Intrusion Detection System Using Hybrid Model of Denoising Autoencoder and Ladder Variational Autoencoder," *2024 10th International Conference on Communication and Signal Processing (ICCSP)*, Melmaruvathur, India, 2024, pp. 214-219, doi: 10.1109/ICCSP60870.2024.10543519.

[NGA24] Next G Alliance, North American 6G Roadmap Priorities, ATIS – NGA, June 2024.[Online]. Available: https://nextgalliance.org/white_papers/north-american-6g-roadmap-priorities/.

[NICT22] NICT, Beyond 5G/6G White Paper, National Institute of Information and Communication Technology, June 2022.[Online]. Available: https://beyond5g.nict.go.jp/images/download/NICT_B5G6G_WhitePaperEN_v2_0.pdf.

[NST+25] H. X. Nguyen, K. Sun, D. To, Q.-T. Vien, and T. A. Le, “Digital Twin for O-RAN Toward 6G,” *IEEE Communications Magazine*, vol. 63, no. 3, pp. 174–181, Mar. 2025.

[Nvi25a] NVIDIA, “NVIDIA AI Aerial Platform,” Accessed on 07/05/2026 [Online]. Available: https://developer.nvidia.com/industries/telecommunications/ai-aerial.

[Nvi25b] NVIDIA, “Aerial cuBB Documentation: CUDA-Accelerated RAN,” Accessed on 07/05/2026 [Online]. Available: https://docs.nvidia.com/aerial/cuda-accelerated-ran/25-1/aerial_cubb/index.html.

[NVIDIA_GPU_26] “NVIDIA H200 GPU”. Available online at: https://www.nvidia.com/en-us/data-center/h200/.

[OKP+24] Seungeun Oh, Jinhyuk Kim, Jihong Park, Seung-Woo Ko, Tony Q. S. Quek2, and Seong-Lyun Kim, “Uncertainty-Aware Hybrid Inference with On-Device Small and Remote Large Language Models”, arxiv.

[O-RAN-XAI] O-RAN Alliance nGRG, “Research Report on Cross-domain AI,” nGRG-RR-2024-02, 2024.

[O-RAN-SBRAN] O-RAN Alliance nGRG, “Research Report on Service-based RAN for 6G Network,” nGRG-RR-2024-11, 2024.

[O-RAN-RANCN] O-RAN Alliance nGRG, “Research Report on RAN-CN Converged Architecture,” nGRG-RR-2024-13, 2024.

[PMS+22] Satheesh K Perepu, Jean P Martins, Ricardo Souza, Kaushik Dey, ‘Multi agent reinforcement learning for intent based service assurance in cellular networks’, Proceedings of IEEE Global Communications Conference, 2022, arXiv 2208.03740.

[QWZ+25] K. Qiao, H. Wang, W. Zhang, D. Yang, Y. Zhang, and N. Zhang, “Resource Allocation for Network Slicing in Open RAN: A Hierarchical Learning Approach,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 11, no. 4, pp. 2584–2600, Aug. 2025.

[Revathy_25] G. Revathy, M. Thangavel, S. Senthilvadivu and M. C. Savithri, "Enabling Smart Cities: AI-Powered Prediction Models for Urban Traffic Optimization," *2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL)*, Bhimdatta, Nepal, 2025, pp. 904-908, doi: 10.1109/ICSADL65848.2025.10933292.

[SBA-RAN2025] “Service-Based Architecture for 6G RAN: A Cloud Native Platform That Provides Radio Capability Exposure and Joint Optimization with CN,” Sensors, 2025.

[SGO+24] N. Souza Neto, M. Gonçalves, D. Oliveira, D. Molinos, R. Moreira, and F. Silva. " Evolved NWDAF Towards a Fully Distributed Artificial Intelligence in the 6G Network Architecture", in Proceedings of the 4th Workshop de Redes 6G, Niterói/RJ, 2024, pp. 15-25, doi: https://doi.org/10.5753/w6g.2024.3378.

[Shafiei2025] F. Shafiei Ardestani, N. Saha, N. Limam, and R. Boutaba, “Towards NWDAF-enabled Analytics and Closed-Loop Automation in 5G Networks,” arXiv:2505.06789, 2025.

[SKT23a] SK Telecom, 5G Lessons Learned, 6G Key Requirements, 6G Network Evolution and 6G Spectrum, White Paper v1.0, Aug. 2023.[Online]. Available: https://newsroom-prd-data.s3.ap-northeast-2.amazonaws.com/wp-content/uploads/2023/11/SKT6G-White-PaperEng_v1.0_clean_20231129.pdf.

[SKT24] SK Telecom, View on Future AI Telco Infrastructure, White Paper v1.0, Oct. 2024.

[SLS+25] Y. Sun, Z. Lin, B. Shi, S. Zhang, S. Ma, P. Jin, Z. Zhong, L. Pan, Y. Guo, and D. Pei, “Interpretable Failure Localization for Microservice Systems Based on Graph Autoencoder,” *ACM Transactions on Software Engineering and Methodology*, vol. 34, no. 2, pp. 1–28, Jan. 2025.

[SNH+25] S. D. A. Shah, Z. Nezami, M. Hafeez, and S. A. R. Zaidi, “The Interplay of AI-and-RAN: Dynamic Resource Allocation for Converged 6G Platform,” in *IEEE INFOCOM 2025 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, May 2025, pp. 1–6.

[STIR/SHAKEN] Sebastian Barros Newsletter, “STIR/SHAKEN for AI Agents? How Telcos Can Move from Identifying Actors to Enforcing Intent”.

[Sundar_25] D. Sheyam Sundar, L. K. Tyagi, S. Kannimuthu, D. Amirtharaj, S. S. Davronboy Ugli and K. A, "Edge AI for Ultra-Fast Personalized Recommendation Systems in E-Commerce," *2025 International Conference on Innovations and Emerging Technologies In AI & Communication*

Systems (IETACS), Mohali, India, 2025, pp. 1148-1153, doi: 10.1109/IETACS68750.2025.11385447.

[TAICS23] TAICS, White Paper on 6G Technology Candidates, TAICS TR-0021(E), v1.0, Dec. 2023. [Online]. Available: https://www.taics.org.tw/eng/Publishing.aspx?PubCat_id=3#.

[TMK+24] T. Tsourdinis, N. Makris, T. Korakis, and S. Fdida, “AI-Driven Network Intrusion Detection and Resource Allocation in Real-World O-RAN 5G Networks,” in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, ser. ACM MobiCom ’24. ACM, Dec. 2024, pp. 1842–1849.

[WLW+23] Q. Wang, Y. Liu, Y. Wang, X. Xiong, J. Zong, J. Wang, and P. Chen, “Resource Allocation Based on Radio Intelligence Controller for Open RAN Toward 6G,” *IEEE Access*, vol. 11, pp. 97 909–97 919, 2023.

[WZS+24] L. Wang, X. Zhang, H. Su, and J. Zhu. “A comprehensive survey of continual learning: Theory, method and application.” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 8, pp. 5362–5383, Aug. 2024.

[XS25] R. Xu, M. Scott, ‘Intent Driven Management enabling highly autonomous networks’, 3GPP SA5 Highlights Issue 10, 2025.

[XSZ+23] L. Xia, Y. Sun, C. Liang, L. Zhang, M. A. Imran, and D. Niyato. “Generative AI for semantic communication: Architecture, challenges, and outlook.” arXiv preprint arXiv:2308.15483, 2023.

[Y23] H. Yang, *Intelligent Robotics and Applications*, 1st ed., ser. Lecture Notes in Computer Science Series, L. Honghai, J. Zou, Z. Yin, L. Liu, G. Yang, X. Ouyang, and Z. Wang, Eds. Singapore: Springer, 2023, no. V.14273.

[YHZ25] Y. Yan, Z. Huang, X. Zhang, Network Automation Enablers in 5GS, May 26, 2025, Network Automation Enablers in 5GS, 3GPP, https://www.3gpp.org/technologies/nae-5gs-ct3#:~:text=First%20published%20December%202024%2C%20in,3GPP%20TS%2029.520%20%5B1%5D.

[ZC22] Y. Zhao and J. Chen. “A survey on differential privacy for unstructured data content.” ACM Computing Surveys (CSUR), vol. 54, no. 10s, pp. 1–28, 2022.

[ZC+23] Zeng, Cheng, et al. "Task-oriented semantic communication over rate splitting enabled wireless control systems for URLLC services." IEEE Transactions on Communications 72.2 (2023): 722-739.

[ZGJ+19] W. Zhao, Y. Gao, T. Ji, X. Wan, F. Ye, and G. Bai, “Deep Temporal Convolutional Networks for Short-Term Traffic Flow Forecasting,” *IEEE Access*, vol. 7, pp. 114 496–114 507, 2019.