Hexa-X: WP4 – Deliverable D4.1

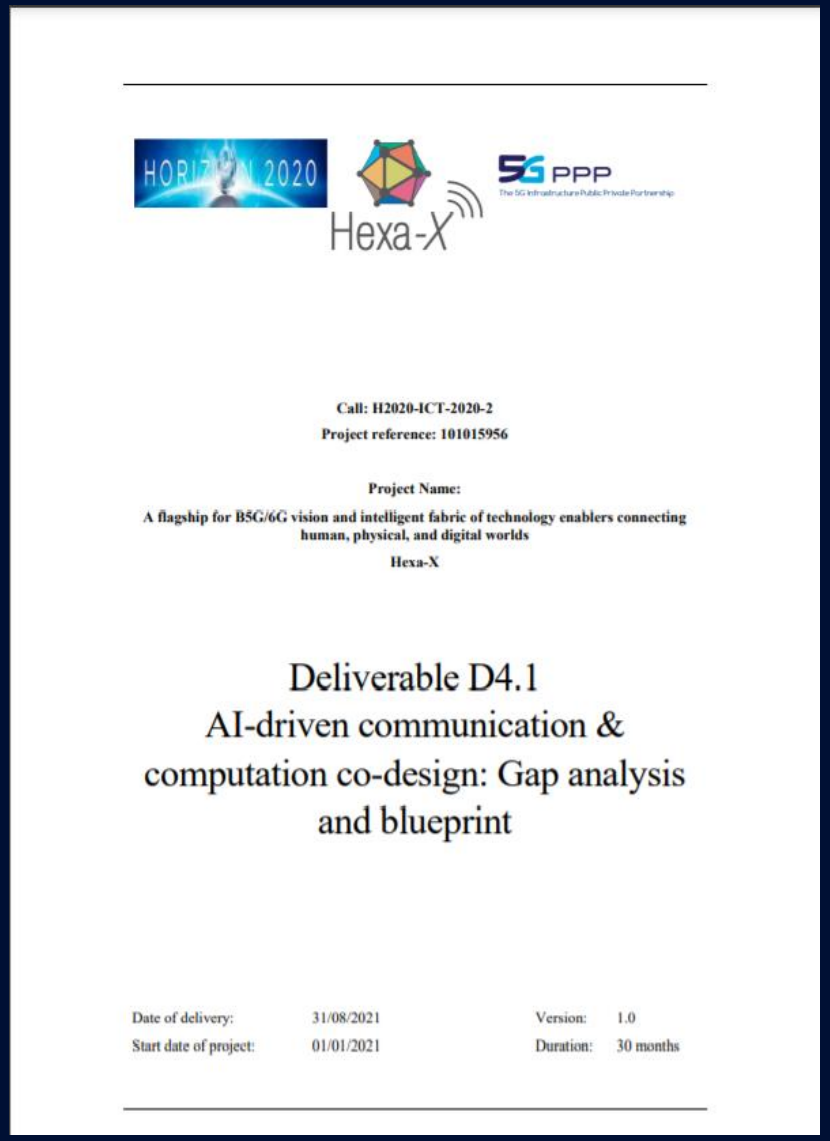
AI-driven communication & computation co-design: Gap analysis and blueprint

31.08.2021

hexa-x.eu

Mission and Scope

- Hexa-X WP4 (*AI-driven communication and computation co-design*) develops concepts for AI-based air-interface design and aims to deliver a secure and sustainable 6G distributed learning platform able to optimally support and address distributed edge workloads and learning/ inferencing mechanisms
- D4.1 provides the rationale leading to the incorporation of AI/ML in 6G networks and documents gaps that need to be addressed to make it possible
- Built upon them, associated problems are detailed and resulting solution directions are presented. Applications in the air interface are considered first and, subsequently, in-network learning methods are investigated
- The deliverable concludes with several architectural recommendations and future work directions



Call: H2020-ICT-2020-2
Project reference: 101015956

Project Name:
A flagship for B5G/6G vision and intelligent fabric of technology enablers connecting human, physical, and digital worlds
Hexa-X

Deliverable D4.1
AI-driven communication & computation co-design: Gap analysis and blueprint

Date of delivery: 31/08/2021 Version: 1.0
Start date of project: 01/01/2021 Duration: 30 months

6G use cases with AI/ML relevance

Use cases & requirements related to AI driven air interface design

Use cases related to AI driven air interface design

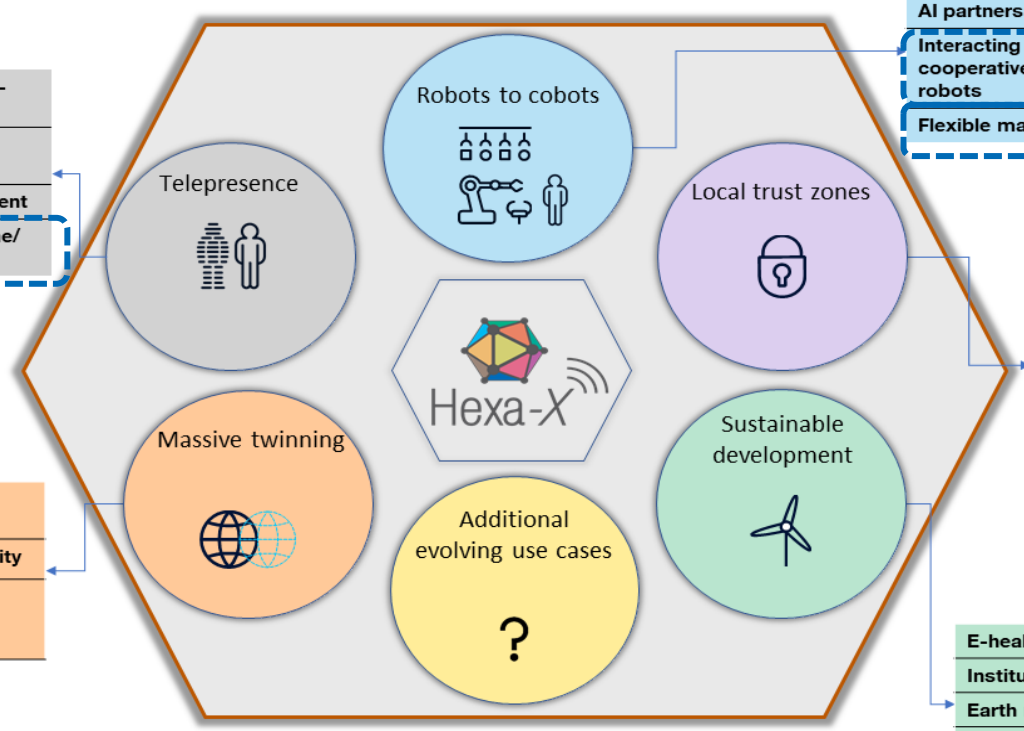
Use case/ scenario title	Brief description (on top of D1.2)	AI/ML-related challenges	Key technical enablers
Merged reality game/ work	Full gaming experience in extended reality.	Needed reliability, bit rate, latency not achieved by means of applying existing communication technologies	AI-driven link level enablers for beamforming design, channel estimation, channel decoding and hardware impairment compensation for improving the bit rate, latency and reliability at link level.
Interacting and cooperative mobile robots	Managing drones/cluster of drones over a 6G network.	Overcoming the challenges with cellular architectures when it comes to managing drone mobility. Need for resource management and mobility management.	D2D communication/cell-free and/or RIS assisted architectures along with AI-driven resource management/link adaptation, AP selection, interference management and mobility management.
Flexible manufacturing	Elevated LiDARs in the infrastructure provide global perception which generates a 3D dynamic map of the factory floor. Communication system utilises the map for optimising directional communications. Also, this map can be used for controlling the robot navigation.	No architecture or signalling for use of such sensor-aided beam management. Also, the communication architecture can be cell free which fits this particular use case well since a set of APs and LiDARs needs to be deployed in the factory floor to deliver seamless communication.	Cell free architectures with AI-driven visual aided beam tracking and beam management, blockage prediction and handovers with mobility.

Focused 6G use cases from AI driven air interface design perspective

Need to improve the bit rate, latency and reliability at the link level using the novel AI-driven link level enablers for beamforming design, channel estimation, channel decoding, hardware impairment compensation etc.

Design of novel D2D/cell free and/or RIS assisted architectures to overcome the challenges of managing mobile robots along with AI-driven system level enablers for resource allocation/link adaptation, interference management, mobility management etc.

Sensing-aided infrastructure providing global perception which generates a 3D dynamic map of the factory floor. The communication system utilises the map for optimising directional communications. A cell-free communication architecture could be utilised to deliver a seamless communication across the factory floor, along with AI-driven enablers such as visual aided beam tracking and beam management, blockage prediction, handovers with mobility etc.



- Fully merged cyber-physical worlds
- Mixed reality co-design
- Immersive sport event
- Merged reality game/ work

- Consumer robots
- AI partners
- Interacting and cooperative mobile robots
- Flexible manufacturing

- Precision healthcare
- Sensor infrastructure web
- 6G IoT micro-networks for smart cities
- Infrastructure-less network extensions and embedded networks
- Local coverage for temporary usage
- Small coverage, low power micro-network in networks for production & manufacturing
- Automatic public security

- Digital Twins for manufacturing
- Immersive smart city
- Digital Twins for sustainable food production

- E-health for all
- Institutional coverage
- Earth monitor
- Autonomous supply chains

KPIs related to AI driven air interface design

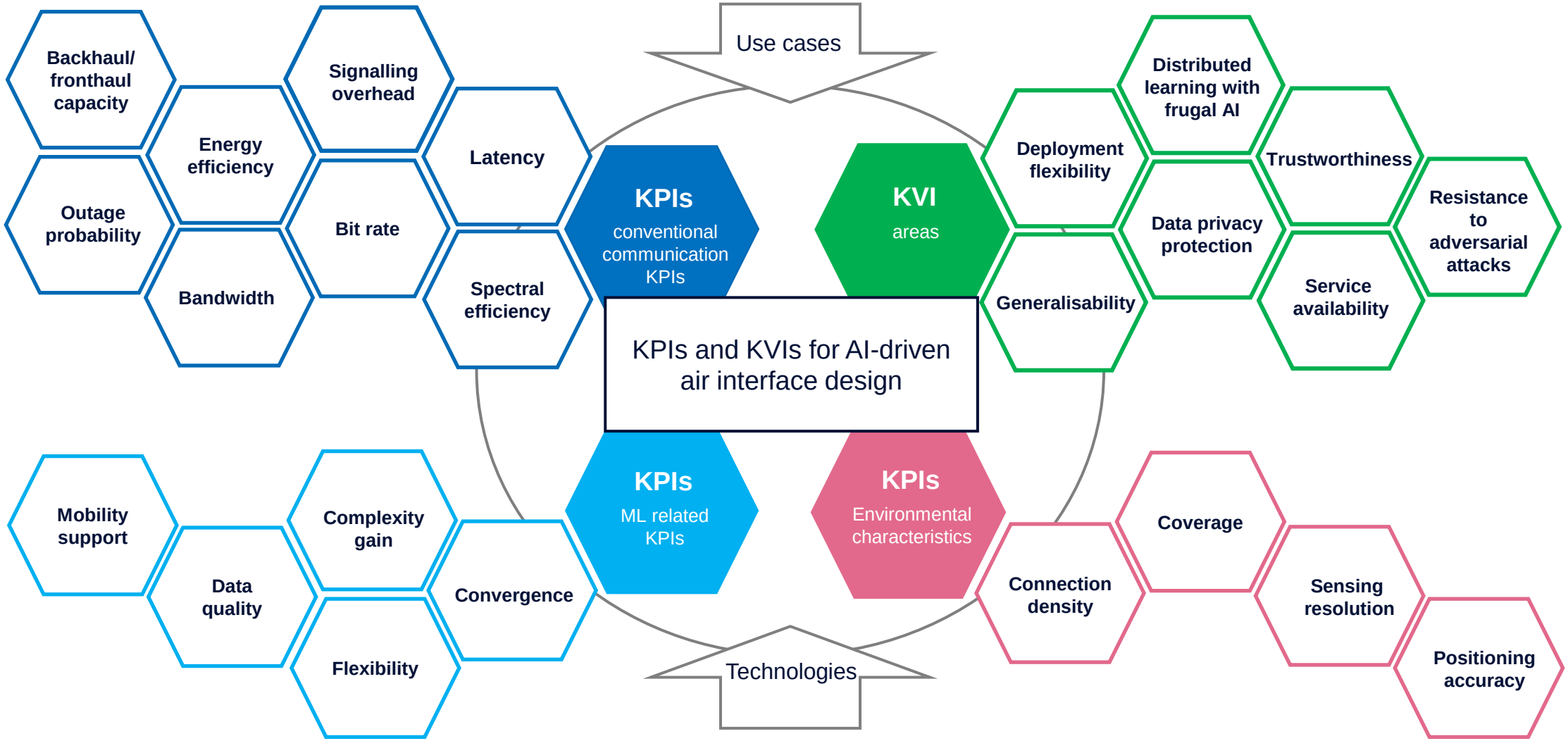
KPI	Brief description	Related KPI area
Latency	Time elapsed between the start and end of air interface functionality design (e.g., channel estimation, decoding etc.).	Conventional communication KPI
Bandwidth	Difference between upper and lower frequencies in a given continuous frequency band.	Conventional communication KPI
Bit rate	Number of bits transmitted per unit time (e.g., seconds).	Conventional communication KPI
Outage probability	Probability that a given information rate is not supported because of variable channel capacity. It is the probability that an outage will occur within a specified time period.	Conventional communication KPI
Energy efficiency	Number of bits that can be sent over a unit of power consumption which is usually quantified by bits per Joule.	Conventional communication KPI
Signalling overhead	Number of radio (e.g., reference/ pilot) signals transmitted for a functionality design to be finalised (e.g., channel estimation).	Conventional communication KPI
Backhaul/frontend capacity	The capacity of intermediate network links (wired or wireless) connecting the core network to the edge of the RAN (backhaul) or central radio controllers to radio heads at RAN edge (fronthaul).	Conventional communication KPI
Spectral efficiency	Information (bit) rate that can be transmitted over a given bandwidth.	Conventional communication KPI
Convergence	Related to training of the ML model. This indicates the loss function value that has been settled with increasing training epochs.	ML related KPI
Flexibility	Ability of the ML model to adapt to different conditions/environments in a timely fashion.	ML related KPI
Data quality	How useful and relevant the data are to model training - assuming the same quantity, higher quality data achieve better model convergence and flexibility.	ML related KPI
Complexity gain	Implementation complexity reduction compared to a non-ML method.	ML related KPI
Mobility support	Ability to support fast moving user connections - relates to flexibility.	ML related KPI
Coverage	The maximum area that can be monitored by the LiDAR sensors.	Environmental characteristic
Sensing resolution	Resolution of the LiDAR sensors.	Environmental characteristic
Connection density	Number of served/connected devices in an area.	Environmental characteristic
Positioning accuracy	Position estimation accuracy.	Environmental characteristic

KVIs related to AI driven air interface design

- **Key Value Indicator (KVI)**: Hexa-X recognises the necessity to expand the fundamental network design paradigm from performance-oriented to both performance- and *value*-oriented in order to fully embrace the 6G vision. Here, value entails intangible yet important human and societal needs such as growth, sustainability, trustworthiness, and inclusion (D1.2).

KVI	Brief description
Generalisability	AI-based models should be able to adapt to unseen scenarios and perform effectively.
Deployment flexibility	Flexibility to deploy same system in multiple scenarios without many modifications to the AI models. Goes hand in hand with generalisability.
Service availability	Ability to perform without any downtime or stability issues (e.g., quick mitigation of adversarial attacks, high model convergence speed).
Distributed learning with frugal AI	Distributed learning enables models to be trained without expensive communication of acquired data. Frugal AI enables learning models based on small amounts of data.
Data privacy protection	Data collection procedures to train the model should adhere to any regulations plus ethical obligations.
Trustworthiness	AI-based models should perform optimally as intended by design without any unauthorised manipulation.
Resistance to adversarial attacks	Capability to perform as intended when faced with adversarial attacks.

Focused KPIs & KVIs from AI driven air interface design perspective



Use cases & requirements related to in-network learning methods and algorithms

Use cases related to in-network learning methods

Use case/ enabling service	Brief description	AI/ML-related challenges	Key technical enablers
AIaaS (enabling service)	ML models and federated explainable AI (XAI) as-a-service; optimal distributed execution capabilities.	ML model task relevance and training level needs to be known; the design of algorithmic strategies for learning.	Edge computing, service-based architecture and Application Programming Interfaces (APIs)
CaaS (enabling service)	Delegating processing tasks to compute nodes; predict capacity and availability.	Inferring power consumption; evaluating future availability and trustworthiness of candidate network nodes.	Network architecture to support service enablement and operation
AI assisted V2X (enabling service)	Digital replica of the real traffic scenario; control and shape massive amount of data.	A wide collection of data distributed on user devices, vehicles and infrastructure elements; Digital Twins of traffic areas used as input by AI algorithms.	Distributed AI algorithms supported by network architecture
Digital twins for manufacturing	Managing infrastructure resources, implementation of different scenarios based on AI/ML predictions.	Transfer of massive amounts of data from the physical to the digital world, reliable / ultra-quick enforcement of decisions.	AI mechanisms with appropriate performance, trust, and greenness levels, for finding and simulating solutions for real world applications
Immersive smart city	Massive twinning for city infrastructure, traffic scenarios, citizen safety; optimised management of control flows	Huge amount of data distributed on millions of user devices, machines infrastructure elements; must be made available for AI applications in a trustworthy, energy and resource efficient way	Distributed AI algorithms supported by network architecture
Interacting and cooperative mobile robots	Flexible production lines require robots and AGVs, UAVs to be adaptive and cooperative	Real-time intelligent decisions based on distributed data; resource efficient data and model sharing; AI/ML for system orchestration	Edge compute, edge AI, distributed AI algorithms
AI partners	AI systems interface with other agents or humans	Decisions based on limited observability and distributed data, interpreting intents	Edge compute, edge AI, distributed AI algorithms
Flexible manufacturing	Flexibility of manufacturing systems through powerful wireless communications with dynamic configuration	Real-time system orchestration of network resources. AI/ML could be used both for optimisation and as a service to enable e.g., dynamic monitoring of the manufacturing process.	Edge computing, edge AI/ML (Compute-as-a-Service)
Merged reality game/ work	Digital interaction of attendees as avatars	Private and secure solutions of ML components	Privacy enhancing technologies in AI

Focused 6G use cases from in-network learning perspective

Addressing network node (device and network side) requests for obtaining ML-based inferencing decisions.

Identify and select available and capable in-network processing nodes for workload addressment

A wide collection of data distributed on user devices, vehicles and infrastructure elements can construct Digital Twins of traffic areas to be used as input by AI algorithms.

AI-as-a-Service (AIaaS)

Compute-as-a-Service (CaaS)

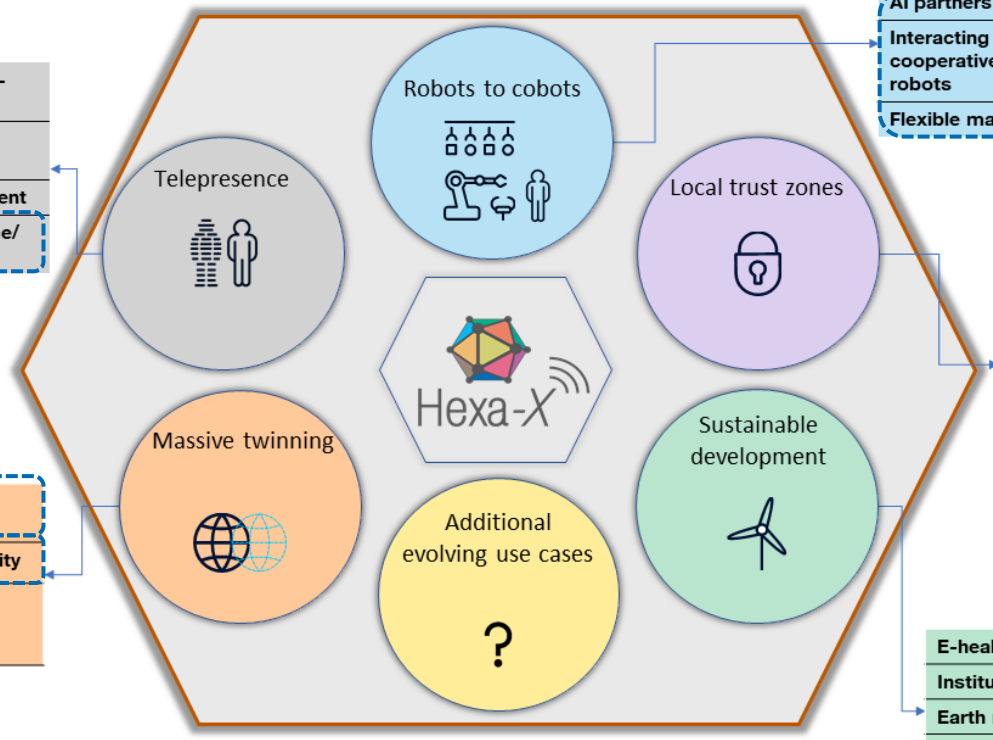
AI-assisted V2X

Private and secure ML-based solutions are needed, satisfying communication and compute dependability requirements.

Fully merged cyber-physical worlds
Mixed reality co-design
Immersive sport event
Merged reality game/ work

Massive amounts of user and machine data shared and processed efficiently by AI models. "What if" scenarios tested at DT level before application in the physical world.

Digital Twins for manufacturing
Immersive smart city
Digital Twins for sustainable food production



Consumer robots
AI partners
Interacting and cooperative mobile robots
Flexible manufacturing

Precision healthcare
Sensor infrastructure web
6G IoT micro-networks for smart cities
Infrastructure-less network extensions and embedded networks
Local coverage for temporary usage
Small coverage, low power micro-network in networks for production & manufacturing
Automatic public security

E-health for all
Institutional coverage
Earth monitor
Autonomous supply chains

Real-time intelligent decisions based on distributed data; AI/ML for system orchestration (e.g. prediction of system disturbances affecting dependability of production).

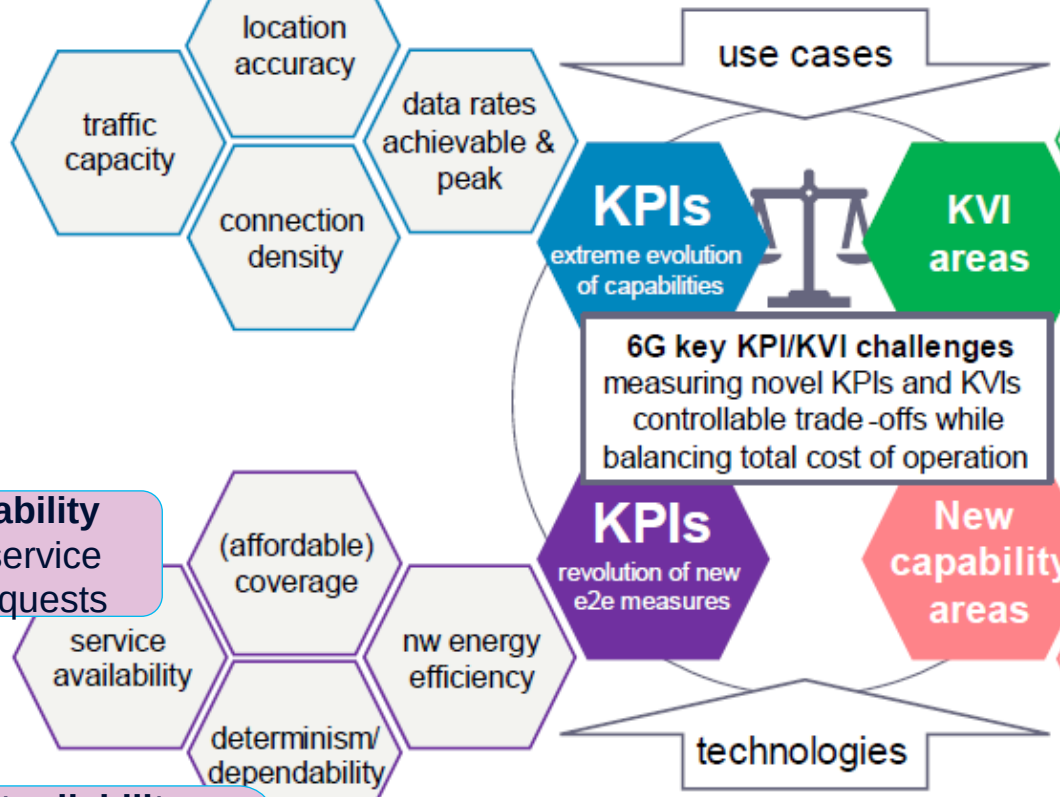
KPIs related to in-network learning methods

KPI	Brief description	Related KPI area
Location accuracy and timeliness	Location estimations enhanced by intelligent fusion with further models (mobility, maps, etc.) and additional data sources - time granularity to be considered jointly with location accuracy.	Extreme evolution of capabilities
AI agent availability	Availability (or readiness) of an AI agent to accept inferencing requests and address them with high accuracy.	Revolution of new E2E measures
AI agent reliability	Capability of an AI agent to accept inferencing requests and provide high accuracy output in a timely manner (within a deadline set by the requesting application).	Revolution of new E2E measures
Latency	AI/ ML components which support (near) real-time decisions also have strict time constraints for inference or training.	New capability areas
AI agent density	Density of devices with AI/ML components considering specific traffic patterns during data sharing.	New capability areas
Interpretability level	Measure of explainability, reasoning, contribution of input factors.	New capability areas
Network energy efficiency	Training/inference optimisation in edge/IoT ecosystem.	New capability areas
Inferencing accuracy	Applicable to many AI functionalities, depends on (and can be traded off for) data volume, inference latency, channel quality in data sharing.	New capability areas

KVIs related to in-network learning methods

AI/ML KVI	Description	Key value areas
Explainability	Ability of the AI/ML agent to provide justification for a recommendation based on model output.	Trustworthiness
Fairness	Ability of the AI/ML agent to perform a decision free from discrimination and bias.	Trustworthiness
Data economy	Capability of achieving high inferencing accuracy with a smaller amount of learning data.	Sustainability (Trustworthiness)
Model complexity	Computational complexity of AI/ML models during either training or inference phases.	Sustainability (Trustworthiness)

Focused KPIs & KVIs from in-network learning perspective (new proposals highlighted)



Location accuracy & timeliness
enhanced by intelligent fusion of multi-source network data

Explainability

Fairness

Data economy

Model complexity

Interpretability level

Inferencing accuracy

AI agent density
impacting traffic patterns of learning/ inferencing data

Network energy efficiency
Training/inference optimisation in edge/IoT ecosystem

Latency
AI/ML supporting real-time decisions

AI agent reliability
accepting inferencing requests and providing high accuracy output in a timely manner

AI agent availability
affected by AI service consumption requests

Research areas & technical enablers

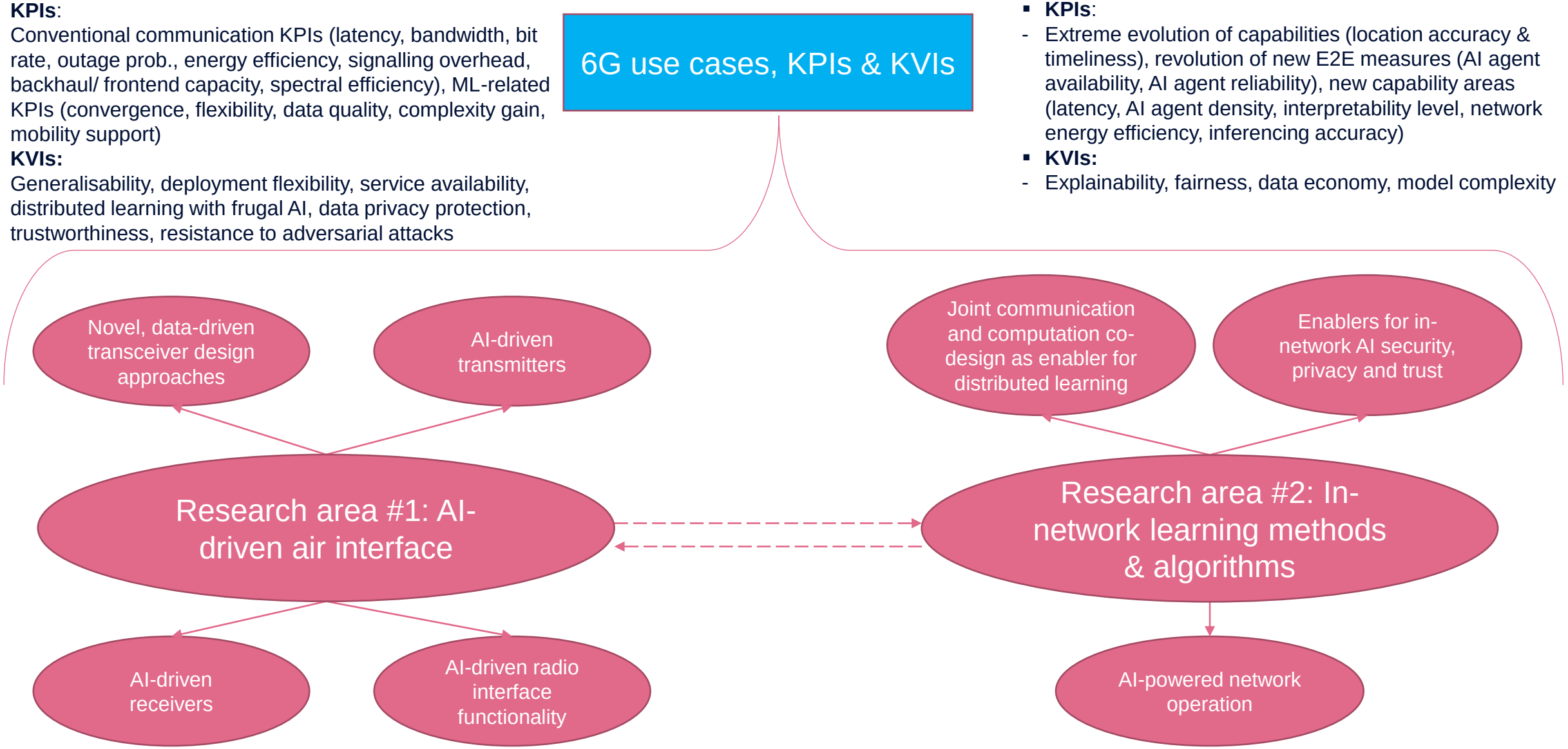
Overview of technical enablers

Initial identified requirements:

- **KPIs**:
  - Conventional communication KPIs (latency, bandwidth, bit rate, outage prob., energy efficiency, signalling overhead, backhaul/ frontend capacity, spectral efficiency), ML-related KPIs (convergence, flexibility, data quality, complexity gain, mobility support)
- **KVIs:**
  - Generalisability, deployment flexibility, service availability, distributed learning with frugal AI, data privacy protection, trustworthiness, resistance to adversarial attacks

Initial identified requirements:

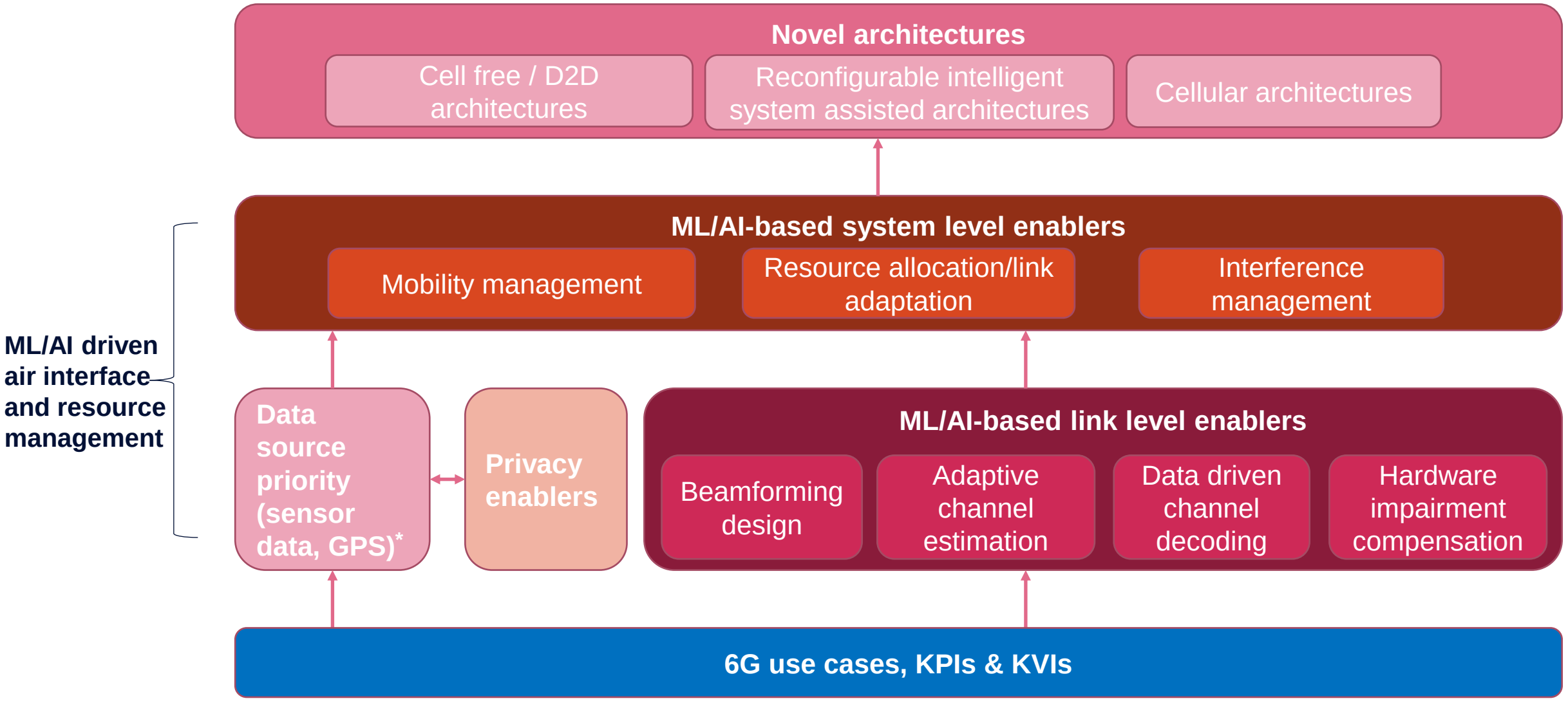
- **KPIs**:
  - Extreme evolution of capabilities (location accuracy & timeliness), revolution of new E2E measures (AI agent availability, AI agent reliability), new capability areas (latency, AI agent density, interpretability level, network energy efficiency, inferencing accuracy)
- **KVIs:**
  - Explainability, fairness, data economy, model complexity



AI-driven air interface design

Motivation & gaps

Overview of proposed enablers for AI driven air interface design



*Data source priority based on availability and privacy requirements

Novel data-driven transceiver design approaches

Topic/ feature of focus	Needs calling for AI/ML-based design
Transceiver hardware impairments	The emerging ML capabilities in communication systems would **enable (1) training models to capture the characteristics of RF hardware impairments**, (2) performing inference based on the trained models **to identify the underlying hardware impairments** during the operation of communication systems, and (3) **optimising the operation of receiver and/or transmitter** algorithms based on the inferred information about the hardware impairments.

AI-driven transmitters

Topic/ feature of focus	Needs calling for AI/ML-based design
Beamforming design, beam management, multi-antenna signal transmission	Innovative beamforming designs and efficient beam management procedures with multi-antennas **need to be developed, as higher frequencies suffer from high propagation loss and scattering**. Beam management procedures become **challenging due to the frequent connection disruptions and fast varying channel conditions**. These channel conditions are further exacerbated when mobility is considered. AI/ML approaches may be applied to address these challenges.
AI for multi-cell, multi-user MIMO	In general, the **optimisation problems in a multi-cell multi-user MIMO system are nonconvex and difficult to solve using the traditional approach based on analytical models**. AI and ML are essential to overcome the limitations of the traditional model-based approach, allowing the future cellular networks to evolve towards more scalable and intelligent architectures.

AI-driven receivers

Topic/ feature of focus	Needs calling for AI/ML-based design
Adaptive channel estimation/ denoising	Channel models are nothing but **imperfect representations of the channel manifold relying on simplifying assumptions**. To optimise channel estimation using physical models, it is required to **handle uncertainties about the antenna gains, positions or radiation patterns (among other system features)**. AI/ML based channel estimation is expected to relax such model-based approach inefficiencies.
ML-based channel estimation for RIS assisted systems	Reconfigurable Intelligent Surfaces (RIS) contain a large number of elements, thereby **increasing the number of links to be estimated**. The RIS itself is a passive component, such that the channel can only be sensed at a receiver by sounding the channel from a transmitter. AI/ML based approaches can be used to **deal with the link scalability issue**.
Low complexity channel estimation	Precise channel state information (CSI) for transmit precoding and beamforming is crucial. **The statistical optimal solution, minimum mean square error (MMSE) estimator, is in general very difficult to implement even with the linear relaxation, i.e., linear MMSE (LMMSE)**. To deal with such algorithm deficiency hardships, data-centric approaches can be applied instead.
Data-driven channel (de)coding for constrained devices	**Efficient and low complexity codes and associated decoders for intermediate length datagrams** - typically a few hundred bits – **need to be designed and optimized** without exhausting the available spectrum resources nor surpassing the constrained capabilities of IoT devices.
Toward an end-to-end driven receiver design	Considering the complete end-to-end radio link, from the transmitter to the receiver, is expected to help **identify which parts of the system should be learned from data**. Such end-to-end driven receiver design scheme will **identify receiver architecture candidates** that are suitable for a native ML air interface, where also some aspects of the transmitter could be learned from data.

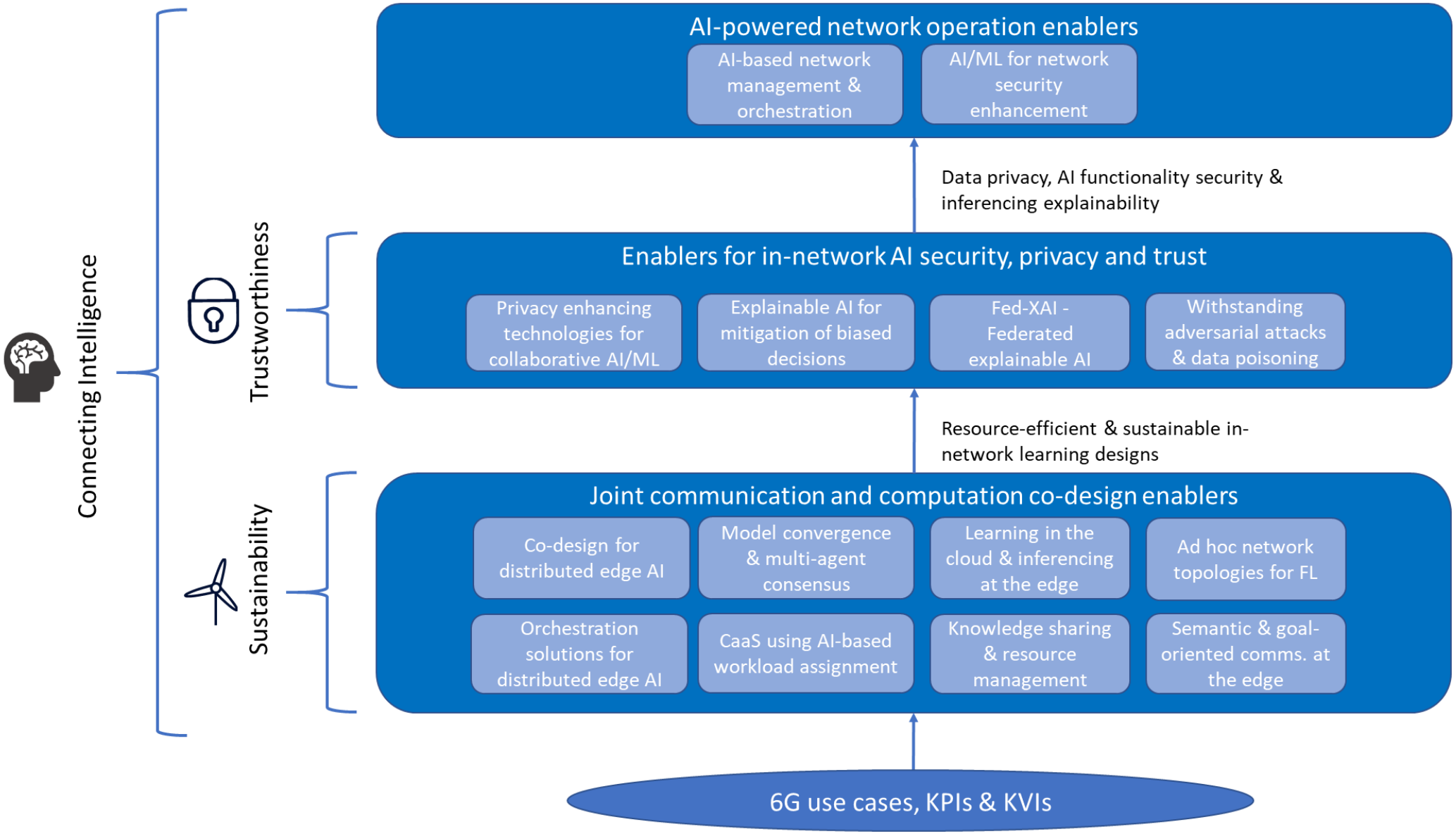
AI-driven radio interface functionality

Topic/ feature of focus	Needs calling for AI/ML-based design
Radio resource management (RRM) based on channel latent variables	Among others (e.g., mitigating model/ algorithm deficiency), a possible by-product of implementing data-centric approaches for air interface design is to **ease RRM using latent variables learned from measured channels**. Such latent variables can be, for example, angles of arrival and path gains. **First envisioned applications are user positioning, and channel mapping/ charting**.
Interference management in cell free massive MIMO	Current pilot allocation/ user-AP association algorithms operate on the assumption that accurate large-scale fading coefficients to all the APs in the system are available. Acquiring this statistic accurately for a large number of APs in a dense deployment is **challenging and may be unrealistic under 5G NR signaling schemes**.
Radio resource allocation for cell-free massive MIMO	State-of-the-art radio resource allocation methods have several challenges, such as **high computational complexity and requiring precise CSI**, resulting in **sub-optimal solutions of complex and non-convex problems, lack of flexibility and parameter sensitivity**, and inaccuracy of the model-based resource allocation methods.
Data importance-aware RRM	An **expected challenge when applying e.g., federated learning in a wireless setup is to prioritise data point transmissions per a data importance criterion**, which is expected to impact the way available radio resources are allocated to wireless devices contributing data. The challenge more specifically lies in the **joint presence of channel uncertainty and data uncertainty** (the latter measured by e.g., the distance from the current model decision boundary).
AI for distributed massive MIMO architectures	The increasing number of antennas in the system heavily **increases the complexity of the optimisation problem to solve**. Distributed massive MIMO poses further implementation issues to become an efficient scalable alternative to existing solutions due to the distributed nature of system components.
Model predictive control (MPC) for MIMO antenna systems	One of the challenges of implementing Reinforcement Learning approaches (RL) for e.g., RRM, is the **lack of quality guarantees, which would be crucial for the service operation**. MPC, the goal of which is to assist complex rule-based systems by learning a close optimal control, aims to close this gap.

In-network learning methods & algorithms

Motivation & gaps

Overview of proposed enablers for in-network learning methods & algorithms



Joint communication & computation co-design as enabler for distributed learning

Topic/ feature of focus	Essentiality to realise a distributed learning platform as part of a 6G system
Communication and computation co-design for improved efficiency of distributed edge AI	Beyond state-of-the-art investigation lies in achieving efficiency improvements for emerging distributed AI architectures. **6G networks should be designed for scalable data sharing to exploit the massive research done in collaborative AI models**, like uplink aggregation, intelligent communication scheduling or localised consensus.
Model convergence and multi-agent consensus in distributed ML	A key gap in decentralised or fully distributed generic ML solutions is **the lack of guarantee of convergence compared to a centralised solution**. Providing such guarantees in a dynamic network environment is a challenging task.
Distributed learning – learning in the cloud and inferencing at the edge	Unlocking the AI/ML potential for computationally and energy constrained devices poses the **challenge of training the models in a way that adapts to environmental changes and considering the individual characteristics of the involved devices**. Distributed learning schemes, where computation-intensive training is processed in the cloud and only small adjustments are handled at the edge devices are of interest.
Ad hoc network topologies for federated learning	Federated learning in a multi-agent wireless network environment prone to failures requires resource replication and network topologies that resist failures and inconsistent views of the entire network. A missing gap consists in the **absence of failure-resilient overlay network structures that take away uncertainties in an abstract network layer**.

Joint communication & computation co-design as enabler for distributed learning (2)

Topic/ feature of focus	Essentiality to realise a distributed learning platform as part of a 6G system
Orchestration solutions for distributed edge AI	**State-of-the-art centralised orchestration approaches exhibit some limitations, mostly related to data privacy, load of traffic to move the data towards the cloud, energy efficiency and latency**. Focusing on orchestration of AI tasks and sharing of AI models, **timeliness is especially critical for actions related to service re-configurations**.
Compute-as-a-Service providing trustworthy and sustainable AI-based workload assignment	Need to support in-network decision making towards discovering and selecting the network node (either a device or a network infrastructure entity such as an edge cloud server) which can **execute a generated processing workload in a high-performance, trustworthy and energy-efficient manner**. **State-of-the-art solution proposals only focus on marginal performance requirements**.
Knowledge sharing and resource management for supporting AI network functionality	The deployment of AI functionality in wireless networks comes at the **additional computational cost of training the relevant models in each participating node, as well as the overhead of exchanging the resulting parameters among them**. Investigation of the optimal resource utilisation and efficient knowledge sharing mechanisms are essential for the operation of AI functionality in wireless communication systems.
Semantic and goal-oriented communication approach for AI/ML at the edge	An efficient training/inference of ML models at the edge requires a **holistic optimisation of communication and computation resources, in order to explore new trade-offs between energy efficiency/resource utilisation, delay and learning/inference accuracy**. The investigation of semantic and goal-oriented communications aims to address these trade-offs.

Enablers for in-network AI privacy, security and trust

Topic/ feature of focus	Essentiality to realise a distributed learning platform as part of a 6G system
Privacy enhancing technologies for collaborative AI/ML	In the collaborative AI/ML context, **privacy technologies** consisting differential privacy, multiparty computation, homomorphic encryption and confidential computing can be used in different settings **to protect model training and deployed models from untrusted aggregators such as ML as a service (MaaS) and the server in the federated learning settings**.
Withstanding adversarial and poisoning attacks in network AI	An open problem is that certain generative ML models **can craft systematic (evasion and poisoning) attacks against ML classifiers**. B5G/6G systems, due to the incorporation of AI components, will need advanced mechanisms that **search for vulnerabilities, especially in a distributed and real-time operation context**.
Explainable AI for mitigation of biased decisions	The adoption of explainable AI (XAI) algorithms and techniques could help in **understanding an AI/ML output-based decision's motivation, detecting potential unfairness and identifying its origin**. Such information can then be used for taking actions to actively **mitigate or even remove decision unfairness**.
Fed-XAI: Federated explainable artificial intelligence	Existing AI/ML approaches ignore one or both of the following requirements: i) the **need for preservation of data privacy, while collaboratively training ML models** and ii) the need for **explainability of the models**. To this aim, the Fed-XAI vision is about devising methods and approaches compliant, at the same time, with federated learning and explainable AI paradigms.

AI powered network operation

Topic/ feature of focus	Essentiality to realise a distributed learning platform as part of a 6G system
AI-based management and orchestration for behaviour-driven adaptation	**AI/ML techniques can be used to avoid both under-resource and over-resource provisioning** by triggering **pro-active scaling actions based on predictions** in order to benefit overall architecture performance. Opposite to the typical reactive approach where services are scaled just after a problem is detected, this AI/ML based proactive approach can be performed based on the **early detection of potential risks**, making possible to sort out the problem before it happens.
AI for Network Security: intrusion detection system architecture and detection procedures	Regarding the topic of **probing and storing of network events**, the objective is to define and prototype an Intrusion Detection System (IDS) architecture to acquire, store traffic data and analyse data traffic in real-time, in order to **detect anomalies**. **One challenge is that the quantity of data to be analysed for security purpose is, a priori, intractable**. Thus, Deep Packet Inspection (DPI), which considers an exhaustive data analysis requiring real-time processes, **cannot be massively used, as it is known to slow down the traffic and be virtually inoperative on encrypted data**.

The role of data in AI/ML enabled 6G networks

Challenges for data management, ownership, and privacy

Data quality, quantity, and availability

- There are certain challenges to address with respect to the *quantity* (directly related to *availability*) and *quality* of learning data
  - A first challenge is to obtain high quality learning data in the sense that **shared context** can be built when e.g., training an ML model, however, **without introducing biases**
  - *Data availability* (spatial and temporal) constitutes another challenge that needs to be addressed by future 6G systems, directly affecting **AI agent availability** and, in its turn, **AI service availability**
    - A measure of temporal data availability may be *data freshness* (measured by the age-of-data)
    - For online learning, a crucial task consists in how to obtain a data set by proper time series sampling
    - *Synthetic datasets* may also be proven useful in the presence of new available training data, for bias minimization
    - Suitable *metadata* may need to be included as data attributes, as well
  - Efficient (learning) *data lifecycle management* is also vital to efficient operation of an AI-enabled network. Functionalities include data curation, provenance, labelling, active learning, distributed storage and deletion

The trade-off between data storage, needed communication signalling and sample processing, on one hand, versus model accuracy and timeliness needs to be studied.

Data ownership and monetisation

- Data generated, transferred and processed within a future 6G network may be categorised, depending on their origin (user application or machine) as either:
  - *network data* (sensing, channel, Quality of Service (QoS) and Quality of Experience (QoE) measurements, network quality analytics etc.) or
  - *third party application data*
- Network data can be assumed to be a "commodity" owned by the network operator, while third party application data may be owned by the data creator or by the application provider, as part of a service subscription
  - Such services may be, for example, transparent services for data storage and sharing (as part of e.g., an object recognition application), or specialised services for efficient management of learning data
- Agreements among various stakeholders (e.g., network operators, service providers, edge cloud providers, end users) on data ownership lead to different implications with regards to data monetisation and structure of data-centric economies
- In 6G systems and using the AIaaS principle, new approaches will arise for monetising data. In particular, a user may decide to request the creation of a “model”. Instead of training data, model parameters will be transferred to the model requestor

This approach, thus, leads to a different learning vendor ecosystem, as compared to the existent ones

Data privacy, security, and integrity

- Not every single generated data set can be exploited for training purposes, as there are limitations related to data privacy and security. In Europe, data must comply with the *General Data Protection Regulation (GDPR)* considering the "Ethics guidelines for trustworthy AI"
  - New rights need to be given to the user related to the ownership of data. For example, the user must be given the right to request removal of certain personal data; for an AI-pervasive 6G network, such data may be, e.g., the identity of a device, its location and additional attributes
- Data privacy and security does not only apply to (human) user application consumer data but also to enterprise data of vertical industries, such as factory automation
  - From this perspective, *local network management* plays a key role in satisfying such enterprise requirements - one can observe an interplay between network and privacy-aware software management to enable automation operations for enterprises
  - A challenge exists in the case *personal data* overlaps with *business data* - a typical solution is to perform data disaggregation and anonymisation prior to any processing
- Regarding *data privacy*, it is expected that 6G systems will require a fine-grained differentiation among various levels of data privacy, as compared to a simplistic separation between "trusted"/ "untrusted" domains
- A trade-off between data privacy requirement stringency and data availability, quality and, ultimately, AI/ML generalisation capability and inferencing accuracy would need to be addressed

6G systems should support systematic features and services for data sharing, aggregation, de-personalisation and anonymisation. 6G systems will also require a fine-grained differentiation among various levels of data privacy

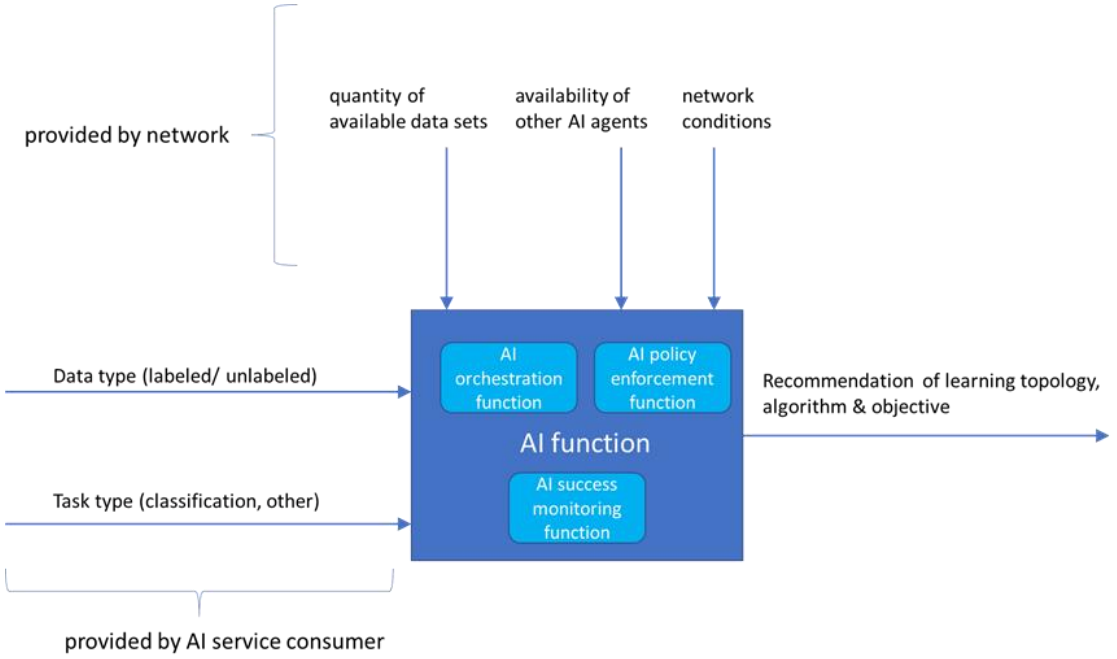
Architectural implications to in-network AI/ML

New architectural components (“AI functions”) containing ML models

New architectural components (“AI functions”)

- In D4.1 some features of new architectural components have been initially described:

AI agent discovery and selection	The challenge consists in how an AI service can perform **discovery of ML model-containing agents** across a network deployment and **select** the one able to issue decision or action recommendations based on inferencing output of highest accuracy and lowest delay and energy consumption. An essential requirement that needs to be addressed by the solution is to **provision an open and generally available AI service with multiple possible instances across a 6G network deployment, each service instance being an "expert" on a specific task to be addressed in a specific part of the network.**
AI service pairing inferencing tasks to learning algorithms and topologies	The challenge is how a service following the AI-as-a-Service (AIaaS) concept can **tailor an incoming inferencing task to the most appropriate learning algorithm and topology**, taking into account the **availability of AI functions, the incurred signaling overhead, the nature of the task itself and the available data.** Solution direction has **implications to network architecture** from the standpoint of **introducing AI functions of given interconnection capabilities across the network, defining the needed interfaces** and **introducing the required signaling protocols.**



Concept of a “holistic AI function”

Summary & future work

First consolidated recommendations for future work

- Summing up, the following consolidated recommendations have been derived calling for addressment both by WP4 and other Hexa-X WPs, as a whole:

- 6G network architecture **should enable and support "end-to-end learning"**, i.e., learning and optimising the transmitter and receiver jointly in a single process. Beyond state-of the-art direction in WP4 is to accomplish E2E learning with advanced waveforms, modulation schemes and channel coding/decoding schemes
- AI-driven transmission (e.g., beamforming optimisation); future direction consists in extending prior work including **applicability of multi-agent systems to real-world, multi-cell, massive MIMO environments**
- Data importance-aware Radio Resource Management **calling for new data structures and communication protocols (e.g., indicators of learning data significance)** in centralised and federated learning
- Architectural implications **supporting online learning** to maximise adaptability to changes in the radio environment are also important for efficient system design
- Aiming at efficient distributed edge AI, **network architecture should leverage on a toolset for communication-efficient inference**. This toolset includes optimised choice of model split points in feature distributed networks, communication-aware model compression (structured pruning, activation pruning at split points) and task-oriented feature encoding
- The network architecture should also be such that **facilitates model convergence and multi-agent consensus in distributed ML**

First consolidated recommendations for future work (2)

- There is also a need for 6G networks to be **flexible enough in enabling the formation of ad hoc network topologies for FL**, motivated by the large heterogeneity of devices and statistical variability of local datasets.
- In terms of orchestrating distributed AI functionality, **an edge orchestrator should be provided the needed information enabling it to semi-autonomously decide to either adjust or redeploy a more robust and reliable AI model at a specific edge location**, increasing the performance of the distributed intelligence.
- **Architectural entities ("AI functions"), network protocols and data structures are needed to apply the Compute-as-a-Service (CaaS) concept** for trustworthy and sustainable AI-based compute workload assignment.
- **Knowledge-based, semantic and goal-oriented communication should be supported** (i.e., qualitative payloads) for sustainable in-network AI operation. There is architectural impact on supported communication protocols foreseen.
- There are architectural implications when it comes to security, privacy and trustworthiness aspects (both referring to attacks to the AI functionality of the 6G network and AI-based attacks to the overall network functionality). **6G network architecture should be supportive of confidential computing (esp. for collaborative AI); protocols are needed to turn the AI functionality of the network to an explainable one to both client applications and NFs, e.g., by using a rule-based approach**.
- **AI functions** should be equipped with capabilities, such as: (i) **AI agent discovery and selection** and (ii) an **AI service pairing inferencing tasks to learning algorithms and topologies**, based on the available data and the AI agents' availability and inferencing capability on the requested task.

Conclusion

- D4.1 presented the rationale leading to the incorporation of AI/ML in B5G/6G networks and documented **gaps** that need to be addressed to make it possible. Built upon them, **associated problems** were detailed and resulting **solution directions** were presented
- The overall storyline of introducing AI in 6G networks, including the motivating challenges and aspired benefits were presented, followed by the definition of fundamental AI concepts and a summary of common practices
- **6G use cases** of particular relevance to AI/ML driven networking were analysed, together with a set of KPIs and KVIs in connection to these selected use cases
- The **role of data** in AI/ML enabled 6G networks was also elaborated; key challenges for data management, ownership, and privacy were explained, as data will be the “fuel” of ML-based algorithms to be developed in the sequel of the Hexa-X project
- Several **technical enablers** and **potential applications** of them to 6G networks were investigated, starting with applications in the **air interface** and continuing with **in-network learning methods**
- Finally, **implications to the envisioned 6G architecture** were documented, along with a number of **consolidated recommendations** for future work

Thank you!

HEXA-X.EU

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101015956.