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

Document properties:

Document Number:	D4.1
Document Title:	AI-driven communication & computation co-design: Gap analysis and blueprint
Editor(s):	Aspa Skalidi (WIN), Tamas Borsos (EHU), Quentin Lampin (ORA), Miltiadis Filippou (INT), Leonardo Gomes Baltar (INT)
Authors:	Aspa Skalidi (WIN), Tamas Borsos (EHU), Quentin Lampin (ORA), Miltiadis Filippou (INT), Leonardo Gomes Baltar (INT), Alessio Bechini (UPI), Andras Benczur (SZT), Giacomo Bernini (NXW), Emilio Calvanese Strinati (CEA), Panagiotis Demestichas (WIN), Pietro Ducange (UPI), Johan Haraldson (EAB), Insaf Ismath (OUL), Dani Korpi (NOF), Ignacio Labrador (ATO), Giada Landi (NXW), Guillaume Larue (ORA), Luc Le Magoarou (BCO), Dileepa Marasinghe (OUL), Francesco Marcelloni (UPI), Ricardo Marco-Alaez (ATO), Mattia Merluzzi (CEA), Jafar Mohammadi (NOG), Markus Mueck (INT), Stéphane Paquelet (BCO), Pietro Piscione (NXW), András Rácz (EHU), Nuwanthika Rajapaksha (OUL), Nandana Rajatheva (OUL), Vismika Ranasinghe (OUL), Alessandro Renda (UPI), Elif Ustundag Soykan (EBY), Emrah Tomur (EBY)
Contractual Date of Delivery:	31/08/2021
Dissemination level:	PU[1]
Status:	Final
Version:	1.0
File Name:	Hexa-X D4.1_v0.1

Revision History

Revision	Date	Issued by	Description
	18.02.2021	Hexa-X WP4	ToC
	09.04.2021	Hexa-X WP4	Draft for internal review
	13.05.2021	Hexa-X WP4	Draft for external review
	24.06.2021	Hexa-X WP4	Draft for PMT review
	26.07.2021	Hexa-X WP4	D4.1 for GA approval
	30.08.2021	Hexa-X WP4	Final

[1] PU = Public

Abstract

This report will provide the rationale leading to the incorporation of AI/ML in 6G networks and document gaps that need to be addressed to make it possible. Built upon them, associated problems will be detailed and resulting solution directions will be presented. Applications in the air interface will be considered first and, subsequently, in-network learning methods will be investigated.

Keywords

6G, services, Artificial Intelligence, Machine Learning, Connecting Intelligence

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

Executive Summary

This report is the first deliverable of project Hexa-X work package four (WP4) "AI-driven communication and computation co-design", led by task T4.1 - "Gap analysis for AI-driven communication and computation co-design". It focuses on the main gaps related to WP4 work and concentrates on associated problem areas (statements) to be addressed, as well as on the description of solution directions. Work concludes with an outlook on planned next steps.

The purpose of this document is to depict motivations for the utilization of Artificial Intelligence (AI) and, in particular, Machine Learning (ML) mechanisms in 6G systems and identify the major challenges that arise. In addition to performing an extensive gap analysis, it also investigates potential approaches and delivers a set of recommendations for future work. Input is yielded from deliverables: D1.1 "6G Vision, use cases and key societal values", D1.2 "Expanded 6G vision, use cases and societal values – including aspects of sustainability, security and spectrum", D2.1 "Towards Tbps Communications in 6G: Use Cases and Gap Analysis", D6.1 "Gaps, features and enablers for B5G/6G service management and orchestration" and D7.1 "Gap analysis and technical work plan for special-purpose functionality". The resulting guidelines are intended to direct the work in the following tasks of WP4, namely task 4.2 (T4.2 - "AI-driven air interface design") and task 4.3 (T4.3 - "Methods and algorithms for sustainable and secure distributed AI").

The overall storyline of introducing AI in 6G networks, including the motivating challenges and aspired benefits is presented, followed by the definition of fundamental AI concepts and a summary of common practices. The role of data is clarified, including considerations related to data quality, quantity, availability, ownership and monetisation, along with aspects on in-network data privacy, security and integrity. Focusing on a subset of identified 6G use cases and their connected performance and value indicators, as detailed in Hexa-X deliverable D1.2, potential applications are investigated, starting with applications to the air interface and continuing with in-network learning methods.

Regarding AI-based air interface design, four main pillars are considered. The first is about novel, data-driven transceiver design approaches, accounting for hardware impairments in the transmitter and receiver radio frequency (RF) chains. Secondly, AI-driven transmitters are considered with an investigation of AI optimization of beamforming and Reinforcement Learning (RL) methods for fast initial access. Thirdly, AI-driven receiver design is discussed, including exploration of functionalities, such as channel estimation and channel decoding, as well as receiver side processing as a single block (end-to-end optimization). Finally, concerning AI-driven radio interface functionality, this document examines different approaches for radio resource management (RRM), cell-free and distributed massive MIMO (multiple input, multiple output) systems, as well as model predictive control of antenna systems.

The topic of in-network learning is organized in three main parts. The section of joint communication and computation co-design investigates different approaches for distributed learning taking into account aspects of trustworthiness, sustainability, efficiency and resilience. Subsequently, enablers for in-network AI security, privacy and trust are analysed, including privacy concerns, explainability features and prevention and mitigation of in-network AI functionality attacks. Lastly, AI-powered network operation incorporating predictive orchestration for behaviour-driven adaptation along with intrusion detection procedures is presented.

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

3D	Three-dimensional
3GPP	3rd Generation Partnership Project
4D	Four-dimensional
4G	4th Generation mobile wireless communication system
5G	5th Generation mobile wireless communication system
5GC	5G Core
5G-PPP	5G Infrastructure Public Private Partnership
6G	6th Generation mobile wireless communication system
AF	Application Function
AGV	Automated Guided Vehicle
AI	Artificial Intelligence
AIaaS	AI as a Service
AIMDD	Active Implantable Medical Device Directive
AnLF	Analytics Logical Function
AoI	Age of Information
AoT	Age of Task
AP	Access Point
API	Application Programming Interface
APT	Advanced Persistent Threat
AR	Augmented Reality
ARCEP	Autorité de Régulation des Communications Électroniques et des Postes
ATIS	Alliance for Telecommunications Industry Solutions
B5G	Beyond 5G
BCH	Bose-Chaudhuri-Hocquenghem-Codes
BP	Belief Propagation
CaaS	Compute-as-a-Service
CAPEX	Capital Expenditures
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CSI	Channel State Information
CVE	Common Vulnerabilities and Exposures
D2D	Device-to-Device

DCB	Deep Contextual Bandit
DDoS	Distributed Denial of Service
DFL	Decentralised Federated Learning
DMO	Direct Mode Operation
DNN	Deep Neural Network
DP	Differential Privacy
DPI	Deep Packet Inspection
DRL	Deep Reinforcement Learning
DSGD	Distributed Stochastic Gradient Descent
DT	Digital Twin
E2E	End-to-End
eMBB	Enhanced Mobile Broadband
EMF	Electromagnetic Field
ENI	Experiential Network Intelligence
ETSI	European Telecommunications Standards Institute
EU	European Union
eURLLC	Enhanced Ultra-Reliable Low-Latency Communication
FDD	Frequency Division Duplex
FL	Federated Learning
FoV	Field of View
FWA	Fixed Wireless Access
GAN	Generative Adversarial Network
GDP	Gross Domestic Product
GDPR	General Data Protection Regulation
GNSS	Global Navigation Satellite System
GPU	Graphics Processing Unit
HAP	High Altitude Platform
HE	Homomorphic Encryption
IAB	Integrated Access/Backhaul
ICNIRP	International Commission on Non-Ionizing Radiation Protection
ICT	Information and Communication Technology
IDS	Intrusion Detection System
i.i.d.	Independent and identically distributed
IoT	Internet of Things
IoX	Internet of Everything

IRTF	Internet Research Task Force
ISM	Industrial, Scientific and Medical
ISG	Industry Specification Group
ITU	International Telecommunication Union
ITU-T	ITU Telecommunication Standardization Sector
IVDMD	In Vitro Diagnostic Medical Devices
JU	Joint Undertaking
KPI	Key Performance Indicator
KVI	Key Value Indicator
LDPC	Low-Density Parity-Check
LEO	Low Earth Orbit
LiDAR	Light Detection and Ranging
LMMSE	Linear Minimum Mean Square Error
LTE	Long Term Evolution
MAB	Multi-Armed Bandit
MARL	Multi-Agent Reinforcement Learning
MBB	Mobile Broadband
MDAF	Management Data Analytics Function
MDAS	Management Data Analytics Services
MDD	Medical Devices Directive
MEC	Multi-Access Edge Computing
MIMO	Multiple Input Multiple Output
ML	Machine Learning
MPC	Model Predictive Control
MMSE	Minimum Mean Square Error
mMTC	Massive Machine Type Communications
MNO	Mobile Network Operator
MR	Machine Reasoning / Mixed Reality
MTLF	Model Training Logical Function
MTTD	Mean Time To Detection
MTTR	Mean Time To Resolution
Mx	Month x after Project Start
NB IoT	Narrowband Internet of Things
NBP	Neural Belief Propagation
NF	Network Function

NFV	Network Function Virtualization
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
NN	Neural Network
NPU	Neural Processing Unit
NS	Network Service
NSM	Network Security Monitor
NTN	Non-Terrestrial Networks
NR	New Radio
NR RedCap	New Radio Reduced Capability
NWDAF	Network Data Analytics Function
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OPEX	Operating Expenditures
OT	Operational Technology
PC	Parity Check
PCA	Principal Component Analysis
PET	Privacy Enhancing Technology
PHY	Physical Layer
PMSE	Programme Making and Special Events
PoC	Proof of Concept
PPDR	Public Protection and Disaster Relief
QoE	Quality of Experience
QoI	Quality of Immersion
QoS	Quality of Service
R&I	Research and Innovation
RAN	Radio Access Network
RAT	Radio Access Technology
RED	Radio Equipment Directive
ReLU	Rectified Linear Unit
RF	Radio Frequency
RIS	Reconfigurable Intelligent System
RL	Reinforcement Learning
RNN	Recurrent Neural Network
RRM	Radio Resource Management

RTT	Round-Trip Time
RX	Receiver
SDG	Sustainable Development Goal
SDN	Software Defined Networking
SDO	Standards Developing Organization
SDR	Software Defined Radio
SGD	Stochastic Gradient Descent
SIEM	Security Information Event Management
SINR	Signal-to-Interference-plus-Noise Ratio
SLAM	Simultaneous Localization and Mapping
SMF	Session Management Function
SMPC	Secure Multi-Party Computation
SNN	Spiking Neural Network
SNR	Signal-to-Noise Ratio
SNS	Smart Networks and Services
SVM	Support Vector Machine
TCO	Total Cost of Ownership
TEE	Trusted Execution Environment
TFF	TensorFlow Federated
TPU	Tensor Processing Unit
TRL	Technology Readiness Level
TSDCI	Telecommunications Standards Development Society
TSN	Time Sensitive Networking
TTM	Time to Market
UE	User Equipment
UN	United Nations
URLLC	Ultra-Reliable Low-Latency Communication
V2N	Vehicle-to-Network
V2X	Vehicle-to-Everything
VNF	Virtualized Network Function
VR	Virtual Reality
VRU	Vulnerable Road User
WP	Work Package
XAI	Explainable AI
XOR	Exclusive OR

XR	Extended Reality
ZSM	Zero-touch network & Service Management

1 Introduction

Hexa-X is one of the 5G-PPP projects under the EU Horizon 2020 framework (project reference 101015956)[2]. It is a flagship project which develops a Beyond 5G (B5G)/ 6G vision and an intelligent fabric of technology enablers connecting human, physical, and digital worlds. Relevant information has been documented in D1.2 "Expanded 6G vision, use cases and societal values – including aspects of sustainability, security and spectrum".

This report is the first deliverable of work package four (WP4) "AI-driven communication and computation co-design", led by task T4.1 - "Gap analysis for AI-driven communication and computation co-design". It focuses on the main challenges related to introducing AI/ML to networking and concentrates on associated problem areas to be addressed within WP4, as well as on the description of solution directions.

1.1 Objective of the document

The purpose of this document is to describe the motivation to use Artificial Intelligence (AI), including, but not limited to, Machine Learning (ML) mechanisms in 6G systems and identify the resulting challenges. This report aims to perform an extensive analysis of the possibilities emerging with the application of AI to networking, investigates potential approaches and delivers a set of recommendations for future work. Input is yielded from deliverables: D1.1 "6G Vision, use cases and key societal values", D1.2 "Expanded 6G vision, use cases and societal values – including aspects of sustainability, security and spectrum", D2.1 "Towards Tbps Communications in 6G: Use Cases and Gap Analysis", D6.1 "Gaps, features and enablers for B5G/6G service management and orchestration" and D7.1 "Gap analysis and technical work plan for special-purpose functionality". The resulting guidelines are intended to direct the work in the following tasks of WP4, namely task 4.2 (T4.2) "AI-driven air interface design" and task 4.3 (T4.3) "Methods and algorithms for sustainable and secure distributed AI". The outcomes of the work presented in this deliverable are consolidated in Chapter 5.

1.2 Structure of the document

The document is structured in the following way: Chapter 2 presents the motivation of this work, identified 6G use cases and performance metrics of relevance to AI/ML in networks, architectural implications, deficiencies in existing approaches, as well as expected gains of the proposed approaches; challenges related to data management, ownership and privacy are elaborated. General concepts are introduced, and the main technological trends are detailed. Chapter 3 explores AI-based air interface design, in four sections: novel, data-driven transceiver design approaches, AI-driven transmitters, AI-driven receivers, and AI-driven radio interface functionality. Chapter 4 targets key gaps and issues of implementing in-network learning methodologies and algorithms. Specifically, the topics addressed are joint communication and computation co-design as an enabler for distributed learning, enablers for in-network AI security, privacy and trust, and AI-powered network operation. The document concludes and provides overall guidelines in Chapter 5.

[2] https://hexa-x.eu/

2 Connecting Intelligence in 6G - overall trends and challenges

2.1 Artificial Intelligence and Machine Learning in 6G networks

Connecting intelligence is identified as a key driver for 6G networks, which will enable effortless interaction among the human, physical and digital worlds [D1.2]. It will provide ubiquitous, real-time, reliable and trustworthy communication among devices with vastly different requirements and capabilities. Fulfilment of these varying and often extreme requirements calls for the application of novel technologies, including AI and ML solutions for a more capable and sustainable radio network and continuous management of edge and cloud resources. Recent advances in AI research have opened the path for a whole new generation of AI-based services, like self-driving vehicles, natural language processing or industrial automation systems with distributed intelligence, which will require new types of communication and computation services from 6G networks for economically reasonable and sustainable deployments.

The increasing significance of AI/ML in 6G can be attributed to several technology enablers. The increasing and ubiquitous compute capacity allows smart allocation of various AI workloads depending on compute cost, device constraints, or how data intensive these workloads are. As AI/ML gained focus in several industries, the creation and usage of dedicated hardware solutions to support and accelerate both learning and inference has increased. AI systems can model the characteristics of structures that are hard or cannot be represented by an explicit mathematical ruleset by learning from large datasets available. The value of data is already well recognised, and more and more data sources are made accessible internally within systems, as well as externally for business and open use. Intelligence will be distributed and spread across agents in self-governed sub-systems of network functions and applications which will then interact with other sub-systems. Safe and efficient communication support among these entities is an essential enabler for scaling out AI capabilities. The recent expansion in the volume of accessible compute and data resources resulted in several major breakthroughs in the areas of AI algorithms and architectures, which brings an explosion of AI/ML services and applications.

For instance, AI/ML brings new paradigms in 6G communication systems on multiple levels. A representative set of new use case families in 6G has been identified in [D1.2]: sustainable development, massive twinning, tele-presence, robots to collaborative robots (termed after as “cobots”) and local trust zones. Use cases in these families pose extreme requirements in the form of stringent Key Performance Indicators (KPIs), as well as the novel concept of Key Value Indicators (KVIs) representing intangible, yet important, human and societal needs such as growth, sustainability, trustworthiness, and inclusion. These targets call for qualitative improvements for ever-increasing demand in data rates, capacity, energy- and spectrum efficiency, in the form of AI solutions. To meet those requirements, 6G will most certainly rely on more spectrum, e.g., exploring the upper mmWave range (100- 300 GHz) as well as on disruptive new technologies (depending on their exploration progress and shown potential), such as Reconfigurable Intelligent Surfaces (RIS) and ML-aided network operations. This new higher frequency spectrum requires novel approaches on algorithmic side, where AI has the potential to harness otherwise intractable optimisations, high complexity calculations or system and environment modelling. Along the improvement of traditional network KPIs, such as peak data rates and latency, 6G is envisioned to also improve on cell-capacity, sustainability, or Electromagnetic Field (EMF) exposure. Those challenges, and the added complexity of

technologies, such as multi-user MIMO (MU-MIMO) complexify RRM procedures and network supervision.

Overall, this increase in network complexity, as well as the diversification of the KPIs and requirements are compelling reasons to integrate AI/ML throughout the network. Indeed, those techniques have been proven to be efficient solutions to problems that exhibit either a model deficit or an algorithm deficit. A model deficit occurs whenever a task is to be performed in a context that cannot be easily summarized by a formal model, for example speech or image content, or when the model is too complicated for a direct exploitation and requires approximation. In the context of 6G, channel and hardware impairment modelling fall squarely in this definition. An algorithmic deficit arises when the task at hand is well defined but finding an optimal algorithm for a known model is prohibitively hard; for example, the game of Go is well defined but finding an optimal strategy is arduous. In 6G, such problems arise in the beamforming, beam management and optimisation problems (refer to sections 3.2.1, 3.2.2).

Furthermore, ML allows new approaches to the design of networks. Data-driven optimisation of network functional blocks allow them to account for phenomena that classical modelling cannot, such as hardware impairments, for instance. While the traditional functional approach allows for segmenting the overall system into simpler functional blocks that are easier to model and optimise, it may also lead to overall sub-optimal system performances. ML techniques are potent candidates to investigate the design of larger functional blocks, hopefully unlocking significant performance gains (refer to section 3.3.5).

ML is also extremely powerful in realising network modules that should be flexible and adaptive to the changing environments. In these cases, ML-based solutions can find the optimal operating point, even when possibly unforeseen events occur. Deep learning methods, where multiple hidden layers are added for flexibility, have the advantage of learning and extracting basic features solely from input data, but it may also turn the model into a less explainable one, thus, it will be less safe to generalise the solution in a previously unseen environment. Using expert knowledge in the model can make the training convergence faster. Therefore, it is important to address proper trade-offs between using domain knowledge and relying on data-driven modelling.

The adoption of intelligent components in mobile gadgets, IoT devices and other machines are also expected to increase. This trend motivates the need to develop intelligent edge functionalities for an AI-native 6G network, capable of supporting AI systems for both network and external/ third-party applications. The wireless environment, energy efficiency, device capabilities and data handling constraints require 6G networks to function as an efficient distributed AI platform. Intelligent functions will also play an increasingly important role in network operation itself by predicting resource needs, optimising and orchestrating application deployments, and simplifying processes for human supervision, in general.

While the application of intelligent functions will increase efficiency in many areas, these goals must be achieved by respecting human values. According to the Ethics guidelines for trustworthy AI [EC19], the solutions should be lawful, ethical, and robust both from a technical perspective and considering its social environment.

2.2 AI/ML primer

Artificial Intelligence (AI), described by Andrew Moore as the "science and engineering of making computer behave in ways that, until recently, we though that required human intelligence", refers today to a very broad field of disciplines, ranging from object classification in pictures to the pursuit of creating an actual artificial, sentient intelligence. Among those

disciplines, Machine Learning (ML), the "study of computer algorithms that allow computer programs to automatically improve through experience" - Tom Mitchell, Machine Learning, Mc Grawhill 1997 -, has enabled computer programs to surpass human capabilities on specific tasks, including playing Chess or Go and hand-writing recognition. Those impressive results find their roots in a few key technical advances, namely the democratisation of hardware computation accelerators, the development of Deep Learning and Reinforcement Learning techniques and the availability of massive databases to learn on. A general overview on Machine Learning can be found in [Apl20], some of the key concepts are summarized here.

ML techniques are classified based on the problem that they target and the learning technique.

Classification problems require the algorithm to associate its input to one or several discrete target values, often called *labels*. One such example is the identification of transmitters, e.g., Wi-Fi, IEEE 802.15.4 in unlicensed frequency bands.

Regression problems consist in matching a function that outputs continuous values, e.g., mapping the household revenue to the property surface.

Prediction problems involve forecasting the likelihood of outcomes based on historical data, for example predicting the probability that a given customer will churn over the next 30 days.

Decision problems require an algorithm to select a set of best actions provided a context, for example what move to perform at Chess to win the game.

To answer those problems, three main class of ML strategies have been proposed.

Supervised Learning: The goal of such techniques is to learn the mapping from an input x to a desired output y. Those techniques require the knowledge of the expected output, making them suited for problems where data are annotated or labelled. The models' "accuracy" is evaluated as a distance between the output of the learned model to y, which is often referred to as *error* or *loss*. Learning the model parameters is, therefore, an optimisation problem, i.e., minimising the *error*/*loss* function. A popular approach to this optimisation problem is called *Gradient Descent*.

Unsupervised Learning: The goal of such techniques is to learn "relations" or patterns in an unlabelled input set x and provide a representation in an output space of smaller dimensionality y preserving properties on those relations. Therefore, such methods learn the a priori probability distribution $P(x)$ and provide a lower dimensional representation of it. Typical unsupervised approaches include clustering techniques such as the K-mean.

Reinforcement Learning: such techniques learn sequences of actions that an "agent" should perform, given its state and its environment state, to maximize the expectation of reward for those sequences of action. Those techniques include Dynamic Programming methods, Multi-Armed Bandits (MAB), Q-learning, etc.

2.3 AI-driven air interface design

One major objective of this document is to investigate the application of ML techniques to the design of the 6G air interface and RRM.

2.3.1 ML-based modelling at the air interface

Exploiting spectrum above 100 GHz is envisioned to be exploited to meet the requirements of 6G new data-intensive use-cases, such as the ones belonging to the immersive telepresence use case family [D1.2]. Radio transmissions at those frequencies face unprecedented signal impairments, as the surface roughness with current manufacturing tolerances is in the order of magnitude of the

wavelength. Radio channel above current FR2 frequencies (i.e., at 26GHz and above), poses additional challenges as diffraction and blockage impediments are emphasised at those frequencies.

Accurate modelling of the channel, environment and hardware impairments is, therefore, a pre-requisite to the efficient exploitation of the spectrum. ML offers a potent set of methods to learn accurate radio channel representations that can be used to estimate the transmitter-receiver links, as described in sections 3.3.1, 3.3.3 and 3.4.1. Using online learning techniques, those contributions expect to enable the learning of models that capture the dynamics of the radio channel or allow the exploitation of the channel at a lower computation cost than state-of-the-art techniques. ML is also deemed a potential enabler to estimate radio hardware impairments. ML, especially Neural Networks (NNs), allow for data-driven non-linear estimators that will be investigated in section 3.1.1.

2.3.2 ML-based optimisation of the air interface

Advanced MIMO techniques, including massive planar array antenna systems, are candidate technologies to improve the overall network capacity and coverage, which is particularly challenging in higher frequency bands; for example, the free space attenuation loss at 300 GHz is 40 dB higher than at 3 GHz. Possible solutions, such as massive-MIMO techniques increase the radio front-end complexity and require improved beamforming and beam management calling for data-driven designs. Contributions detailed in sections 3.2.1 and 3.2.2 investigate such ML applications.

Novel Radio Access Network (RAN) architectures are also envisioned for 6G, including cell-free or cell-less distributed architectures. Those user-centric architectures raise the complexity of the network RRM, allowing unprecedented allocation of pilots and Access Points (APs) to User Equipment (UE). This paradigm change challenges existing radio resource allocation schemes and advocates for optimisation strategies of reduced implementation complexity that account for multiple complex requirements and constraints expressed in the form of maximum tolerable intercell interference, power limitations, and fairness. Data-driven approaches to those optimisation problems, as well as their challenges are investigated in section 3.4.

ML at the air interface is also under investigation in 6G: as specific components in combination with traditional expert algorithms in a functional approach, as studied in section 3.3.4 or by means of a systemic approach (section 3.3.5). In a functional approach, the communication chain is a composition of blocks with well identified functions that are optimised individually based on local metrics. On the opposite end, the systemic (or, End-to-End - E2E) approach proposes to tackle the overarching design problem of the transmission chain either as a single task, or as a few tasks, where each task incorporates multiple transmission chain blocks. The key motivation of this approach lies in the potential performance gain of jointly optimising multiple subtasks. Those two approaches are most often related to as “clear-box” and “black-box” approaches. In the “clear-box” setting, models are defined based upon standard signal processing algorithms and expert knowledge leading to generally rather low-complexity and explainable models. On the contrary, in the “black-box” approach, the models are defined following more disruptive deep learning approaches with often more complex, but also more expressive and potentially better performing models.

2.3.3 ML at the air interface - the key challenges

Using ML at the air interface introduces a paradigm shift in its design. Per the traditional approach, algorithmic solutions are devised based on an a-priori model of the problem at hand, for example channel coding could assume a binary erasure channel. Most often, this a-priori

modelling allows for an analytical study of the performance of algorithms, including performance bounds. On the contrary, ML techniques are data-based, and the algorithmic solution is tightly bound to training datasets: there is often no guarantee that a trained model can deliver expected performances in a controlled setup and no proof of correctness other than its actual performance. Those observations highlight the need of explainable ML components, whose behaviour can be analysed and explained.

Training and validation datasets are also of utmost importance to ensure models perform as expected in a live network. Being a key enabler of ML performance, data used for training must, therefore, accurately represent real-world radio signal conditions in which models will then be executed. This raises the question of the "quality" of the data and advocates for metrics to evaluate it, for example its freshness or its correlation to the rest of the dataset. This also highlights the need for data curation and/or selection procedures. Caution is especially required when offline training is in use, as models' generalisation capability is an absolute requirement. Online automatic training raises other equally challenging issues, such as the definition of KPIs to monitor the training and inference routines and metrics to evaluate the relevance (or, environmental representativeness) of data used for training.

Energy-efficiency and hardware complexity is another key challenge. Low-power hardware supporting NNs often provide a limited set of efficient computation operators, e.g., matrix multiplications and simple activation function such as the Rectified Linear Unit (ReLU), and small datatypes (e.g., unsigned integers of size of 8 or 16 bits). Building models according to those specifications is key to their efficiency. Otherwise, key computation operators specific to the domain of communication systems must be identified to be implemented in future NN dedicated hardware.

2.4 In-network learning methods and algorithms

Another major objective of this deliverable is to highlight the role of joint communication and computation co-design, which is needed to realise a distributed learning platform as part of a 6G system useful to both network operation and end-user applications. The incurred research directions calling for novel solutions refer to the concept of "edge AI" incorporating collaborative multi-agent architectures, ML model decomposition and data parallelism principles (refer to section 4.1.1 of the deliverable); the impact of implementing these architectures on network resource allocation needs to be clarified beyond state-of-the-art status. Furthermore, the challenge of ML model training convergence and consensus among multiple AI agents in distributed learning setups - having a centralized architecture as a benchmark - needs to be addressed, considering the heterogeneity of network conditions and the existence of learning "strugglers", i.e., network devices (end user or network infrastructure ones) having insufficient compute, memory and/or storage resources to update a local model in a timely fashion per e.g., a learning synchronisation requirement; this topic is elaborated in section 4.1.2. Additionally, the potential of a setup involving learning and inferencing at different "depths" of the network (e.g., edge versus cloud) can be thought of, where the transfer learning and knowledge distillation paradigms may be proven useful (section 4.1.3). Moreover, as part of this overall research direction, ad hoc network topologies need to be explored for geo- and network capability-aware Federated Learning (FL) aiming at enhancing robustness to failures (section 4.1.4). A further challenge consists in orchestrating the instantiation of AI agents aiming to increase the performance of distributed intelligence (section 4.1.5). In addition, the concept of providing Compute-as-a-Service (CaaS) [D1.2], aiming at serving devices of limited capabilities - but not only limited to them - needs to be dealt with aiming at sustainable and secure workload addressment (section 4.1.6). It can be also argued that AI network functionality can be supported by e.g., the use of

Blockchain technology, knowledge sharing techniques and resource management (section 4.1.7). Finally, semantic and goal-oriented communications for edge AI/ML can be exploited in terms of addressing trade-offs between energy efficiency/resource utilisation, delay and learning/inference accuracy (section 4.1.8).

Apart from efficiently and sustainably allocating network and computing resources, another challenge of paramount importance is to ensure the secure, data privacy-keeping and trustworthy AI functionality of a 6G system with pervasive AI capability (section 4.2 of the deliverable). These aspects are crucial, as they have both legal and regulatory implications; for example, the General Data Protection Regulation (GDPR) - [GDPR]. Furthermore, privacy of data and security of learning datasets and AI functionality as a whole, whether put into practice to support home, work or other human or machine-related applications, have certain ethical consequences, e.g., on fairness of inference-based decision making and its aftermath. In the sequel of this deliverable, a number of directions will be elaborated starting from the area of Privacy Enhancing Technologies (PETs) for collaborative in-network AI/ML functionality (section 4.2.1). Further, as data is the driving force for efficient AI/ML-based system operation, since learning datasets (and, even the parameters of an ML model) are prone to deliberate modification (also termed as "poisoning") and stealing, adversarial attacks to AI network functionality must be effectively mitigated (section 4.2.2). Key issues in AI trustworthiness will be also provided, along with possible solution directions, i.e., (i) expanding the network's data analytics functionality by identifying and mitigating unfair/ biased ML-based decisions (section 4.2.3) and (ii) turning a FL setup to an explainable one, so as to improve decision transparency, while not sacrificing privacy of data at the same time (section 4.2.4).

The pervasive AI functionality across the network - when effectively supported, leading to a communication / computation resource-efficient platform which is also energy-efficient, secure, and trustworthy - can be exploited for support or even complete undertaking of network operation functionalities (section 4.3). The broader vision of such functionality undertaking is to automate network operation, thereby introducing a level of autonomy to the future 6G system (please also refer to [D6.1]). The advantage of such automated operation is relaxing the need for human-driven monitoring and network adaptations, as a reaction to encountered triggers, e.g., unexpected increase in user loads, deterioration of signal quality at certain locations etc. To achieve that, AI/ML network functionality can enable predictive -instead of reactive- orchestration of resources; as a result of such proactivity, the earlier a problem can be identified, the timelier it will be resolved, thus, improving performance and limiting risks and incurred costs (section 4.3.1). On top of predictive orchestration, in-network AI functionality can be harnessed to detect security attacks to a wireless communications system by means of detecting abnormalities in databases stored across the network (section 4.3.2).

2.5 Architectural implications to in-network AI/ML

Regarding the design of a network architecture capable of supporting AI/ML functionality in a 6G network, which aims to offer the abovementioned performance and value advantages, AI agents carrying (trained) ML models can be instantiated within AI "functions" allocated across the network. Today's paradigm of the 3GPP specified Network Data Analytics Function (NWDAF) is a step towards this direction, however, to enable different inferencing topologies that may be needed depending on data availability, it would need to generalise in a way that AI functions can be instantiated and accessed anywhere in the network by proper protocols. Moreover, as there is different availability of data across the network, different network conditions and AI agents of different knowledge levels focusing on heterogeneous inferencing tasks, an AI service would need to be cognisant of these conditions and recommend the most

appropriate learning topology and algorithm. The topics of: (i) AI agent discovery and selection and (ii) AI service pairing inferencing tasks to learning algorithms and topologies are elaborated in sections 2.5.1 and 2.5.2 below.

2.5.1 AI agent discovery and selection

To perform both AI-based air interface design achieving extreme performance and implement in-network learning algorithms aiming at automated, AI-powered network operation, specific network architecture components are needed to facilitate discovery and selection of AI agents (e.g., containing ML models) across a network deployment. Performance criteria for such AI agent selection are the issuing of inferencing output of highest accuracy and lowest delay and energy consumption. The entity in need of consuming an AI service can be e.g., a Network Function (NF) for the needs of network automated operation or any (for instance, vertical-related) user application, while the service producer can be e.g., an AI task orchestration entity.

In terms of state-of-the-art, [D1.2] outlines the concept of AI-as-a-Service (AIaaS), where the ultimate goal is to provide a reasoning service for e.g., prediction, anomaly detection, classification, tactical decision making or other tasks. The service may rely on network AI capability, but also on non-AI related statistical algorithms, such as Autoregressive Moving Average (ARMA) models [WiKi+21]. In [STG+20], the authors provide an overview of the 3GPP specification TS 29.520 defining the NWDAF [29.520], and some simulation results comparing NN algorithms with linear regression to predict the network load. [Dua21] surveys the latest developments in the standardisation of network management architectures carried out by 3GPP, ETSI, and ITU-T and analyses how cloud-native network design may address network management challenges. However, this work rather focuses on AI/ML-assisted network management and does not discuss the wider need for an AI agent (that may or may not be instantiated at the network infrastructure side) to be discovered by AIaaS consumers or other AI agents for the purpose of knowledge sharing, relating to tasks other than network management and automation.

Going beyond state-of-the-art, an essential requirement that needs to be addressed by solutions 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 - of course, it would also be possible for the "range" of an AI service instant to be wider than a local deployment area. Further, it should be ensured that a given AI service instance is available in time and reachable by any network entity (either at network infrastructure or even instantiated at user device side) within its "range" of operation, provided that it is first authenticated and authorised to provide data and/or inferencing requests. In this context, some of the functional requirements that need to be addressed by future solutions are the following:

- A general interface providing access to AI resources external to the service calling entity needs to be defined.
- A mechanism for discovery and selection of available AI agents. An AI agent can be split into two independent functional entities, one focused on model/data management and another one concentrating on inferencing tasks. Implementation of the mechanism requires e.g., a frequently updated repository of available models (ML-based or others) instantiated at available AI agents/ AI functions.
- The selection of an AI function (equiv. AI agent) responsible for producing inferencing output needs to take into account several aspects and possible trade-offs between e.g., relevance of the trained ML model to the type of the requested inferencing task, the

incurred latency in obtaining and dispatching inferencing output to the AI service consumer and the energy footprint of operation.

After an AI orchestration function performs discovery and subsequent selection of the appropriate AI function (or, AI resources, in general), the AI service consumer (i.e., network entity or user device) will attach to the selected AI function and request for inferencing output. The request can be optionally accompanied with additional training data, as depicted in Figure 2-1.

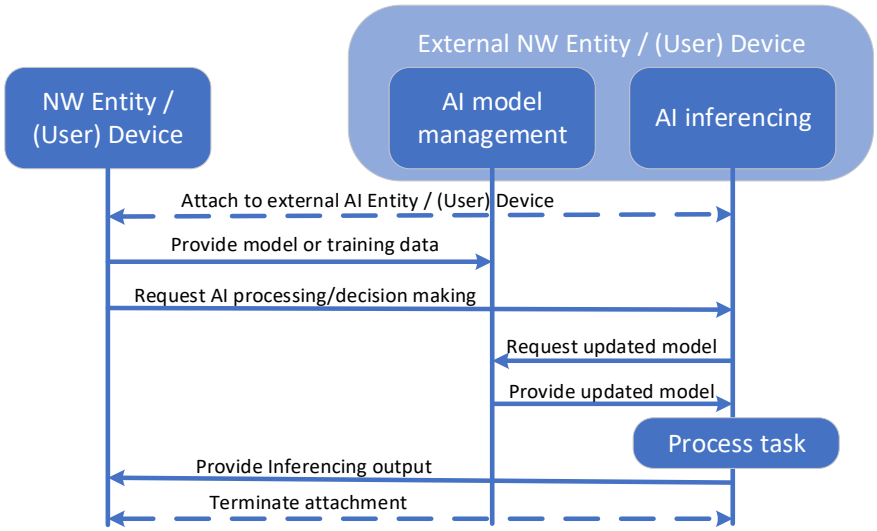

Figure 2-1: Signal flow chart illustrating AI service (training and inferencing) consumption per the AIaaS concept (after AI function discovery selection)

In Figure 2-2, the result of the procedure is illustrated when considering a collaborative learning setup, where, both the AIaaS consumer and the selected AI agent are capable of training their local models. Aggregation of these models is performed either at the AI service consumer side or at the selected AI agent side.

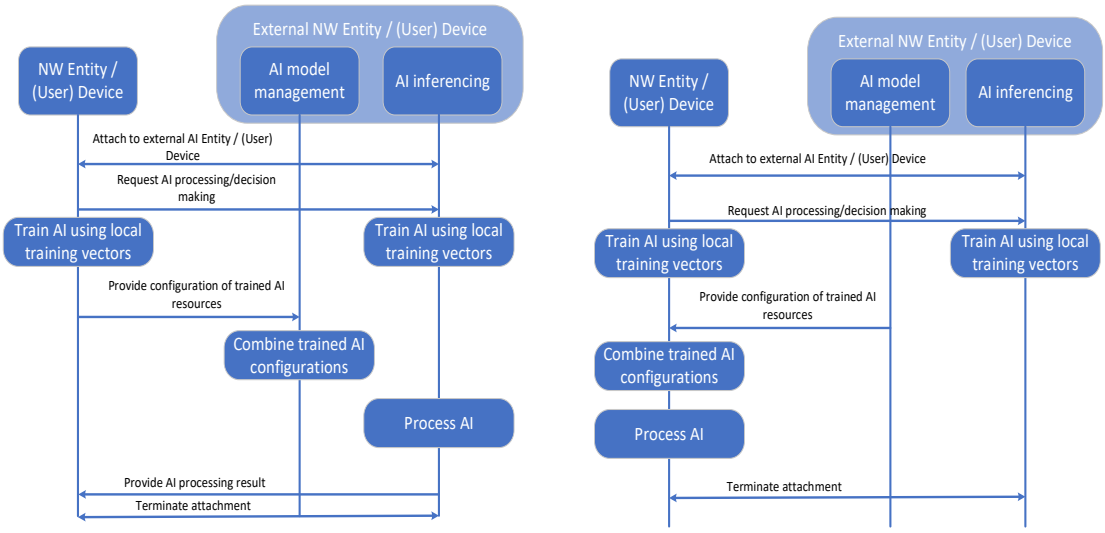

Figure 2-2: Signal flow chart illustrating how AIaaS applies to collaborative learning - inferencing is performed at the AI service consumer (left) or at the selected AI agent/ AI function (right)

2.5.2 AI service pairing inferencing tasks to learning algorithms and topologies

In AI/ML systems, there can exist different learning algorithm and architecture taxonomies. Categorisation is based on different criteria, such as: (i) the structure of the available training data sets (e.g., labelled versus unlabelled data), (ii) the AI function/ agent deployment (i.e., single-agent versus multi-agent systems), (iii) the level of distribution when it comes to multiple-agent scenarios (i.e., centralised, distributed learning) and the objective of the learning framework (i.e., aiming an overall deployment-wide model for maximum generalisation capability or exploiting knowledge-level variations of multiple AI agents for data frugality improvement). In other words, the problem is how a service following the AIaaS concept, as described in [D1.2], can tailor an incoming inferencing task to the most appropriate learning algorithm and topology, taking into account the availability of AI functions, the incurred signalling overhead, the nature of the task itself and the available data.

There are numerous works focusing on different learning architectures and topologies. They include centralised, federated [MMR+17] [KMR+16], and distributed learning [VWK+20]. Centralised learning is typically used as a coordination benchmark, however, it suffers from communication inefficiency and security attacks, as it involves a single point of failure, For distributed learning, examples refer to transfer learning [WLT+21] [PPK+21], multi-task learning [RPB+20] and meta-learning (or, "learning to learn") [HZY+20]. However, considering the above mentioned categorisations, a solution is currently missing to the problem of a service jointly selecting the proper learning type (i.e., supervised/ unsupervised learning, Reinforcement Learning - RL), learning structure/ topology and/or learning model (e.g., linear/ logistic regression, NN, Support Vector Machine - SVM etc.) for in-network inferencing. Some works, for example, [MCC+19] rather investigate marginal aspects, such as topology selection for deep learning.

As shown briefly in Figure 2-3, a solution direction can be the providing of a *"holistic" AI service* to the calling service consumer, via an AI function taking into account a number of factors before launching a learning procedure (i.e., the task initialisation or the obtaining of a model update) and subsequently performing inferencing and dispatching of the AI system output to the AI service consumer. Of course, this 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 signalling protocols.

To provide such a holistic AI service, the service provider (AI function) may provide the following functionalities, altogether as part of a single entity, or in separate entities: (i) an *AI orchestration function* for AI agent discovery and selection; (ii) an *AI policy enforcer* to implement the recommended policy, and (iii) optionally, an *AI success monitoring function* performing inferencing accuracy, communication efficiency, and security monitoring. The needed interfaces and communication protocols will be studied in Hexa-X project, in collaboration with WP5 - "Architectural enablers for B5G/6G".

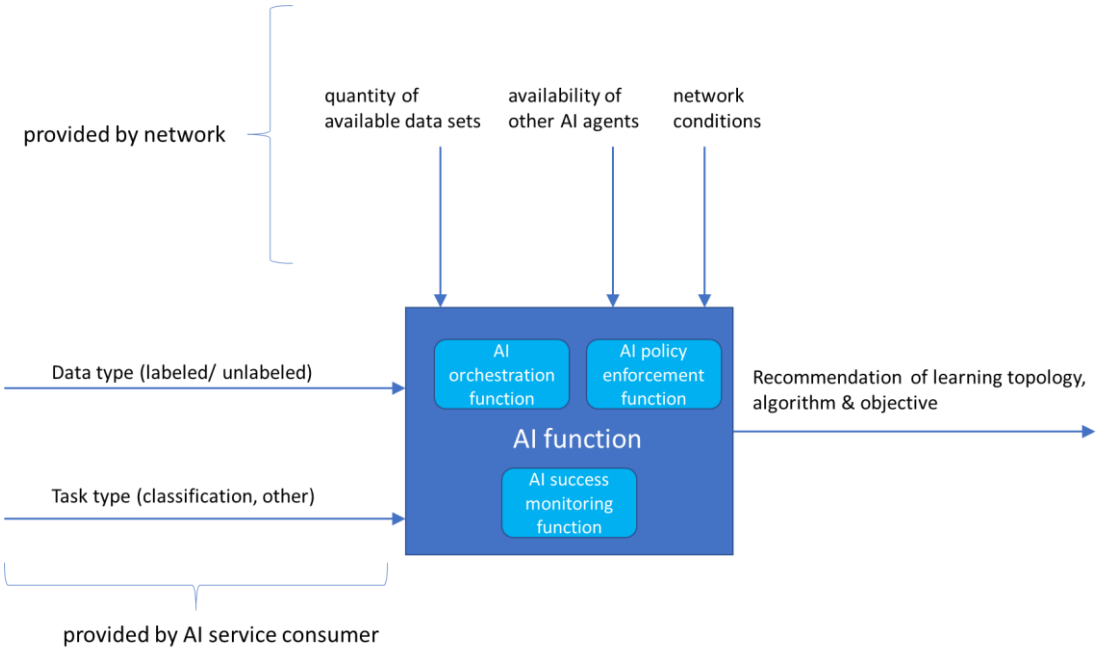

Figure 2-3: Principle of a holistic AI function

2.6 Use cases with AI/ML relevance

Use cases and KPIs/KVIs are analysed from a WP4 perspective and with a focus on technical enablers in this document, based on their initial definition and refinement in D1.1 and D1.2 [D1.2]. As part of the ongoing work within Hexa-X, harmonisation and alignment of initial use cases and KPIs/KVIs of D1.1 and D1.2 with findings from technical work packages will be performed in WP1 and outcome will be available in D1.3, due in February 2022.

2.6.1 Use cases, KPIs and KVIs relevant to AI-driven air interface design

As elaborated in section 2.3, there are several potential benefits when it comes to utilising AI/ML based techniques for air interface design, ranging from more energy, complexity-efficient and radio environment adaptable signal transmission and reception to more efficient RRM, especially regarding large scale optimisations. Such benefits may be useful to address the ever-increasing performance requirements envisioned for 6G communications, judging from the 6G use case families, as introduced in [D1.2]. In Table 1 below, three representative 6G use cases have been selected as ones which can be more profoundly addressed by data-driven design of the incurred air interface. Table 1 provides further characteristics of these identified use cases (as compared to their descriptions in [D1.2]), their incurred challenges from an AI/ML perspective and key technical directions (or, enablers) worth investigating towards addressing these use cases.

Table 1: 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.

The selected use cases appearing in Table 1 are accompanied by a set of KPIs that need to be satisfied, which are listed in Table 2. These KPIs can be categorised into "conventional" communication KPIs -as known from past cellular communication generations- and ML related KPIs (as ML-based air interface design is the focus of this deliverable). At the end of the table some environmental characteristics that need to be taken into account by AI-driven air interface design approaches are detailed. The targeted values of the mentioned KPIs will be proposed as part of future Hexa-X work.

Table 2: 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

Table **3** lists down the identified Key Value Indicators (KVIs) of high importance to AI-driven air interface design.

Table 3: KVIs related to AI driven air interface design

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 (resistance to adversarial attacks).
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.

2.6.2 Use cases, KPIs and KVIs relevant to in-network learning methods and algorithms

Many of the future 6G network use cases will be intertwined with various forms of learning and intelligence involving both applications and network functions. Apart from air interface design, the expected use of AI/ML technology in 6G will impact data management, optimality of compute and processing functions, network automation, service availability, while it will call for the implementation of suitable mechanisms for trustworthy operation. Some of the 6G use cases relevant for Hexa-X WP4 are, therefore, primarily connected to the more general enabling services in 6G networks (see [Section 4.2.7, D1.2]), which are useful to address requirements of multiple vertical industries. Besides these generic services, there are some use cases, for the addressing of which, AI capabilities are particularly important; these use cases (also selected from [D1.2]) are also highlighted in this section. Use cases and more generic enabling services with high relevance to WP4 work appear in Table **4**.

Table 4: 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 (see note)	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 (see note)	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

Among the enabling services AI and Compute as a Service, as well as AI-assisted-V2X are significantly impacted by AI/ML solutions, offering ML models with capabilities such as efficient execution, explainability and privacy, but also providing in-network solutions for assessing important execution features for candidate network nodes like power consumption, future availability or trustworthiness. In relation to massive twinning use cases, a recent proposal in Internet Research Task Force (IRTF) Network Management Research Group [NMRG] considers the applicability of DT network platforms, leveraging AI models to help the network admins in assessing specific behaviours and run "what-if" scenarios that could not be tested and evaluated easily in the physical network. In particular, this approach advocates to apply fuzzing testing techniques on the twin environment, with interactions and conditions similar to the production network, permitting to identify and solve vulnerabilities, bugs and zero-days attacks. This is especially relevant if we consider the applicability of a DT network platform to generate datasets for ML, providing a realistic environment including network topologies. These datasets are generated in a controlled context, allowing information access by third parties.

The Hexa-X vision of 6G is built on important human and societal key value areas, such as sustainability, trustworthiness, and inclusion [D1.2]. Besides these value areas, flexibility is also seen as a core multifaceted network capability, which enables novel application areas, use cases, as well as easier network deployment and operations. AI/ML as a technology can contribute to progress in the identified key value areas by taking into account key aspects during the design and application of the developed algorithms. These aspects lead to definition of key values and KVIs within the domain of AI/ML, which drive research directions and outline more specific key performance metrics. The identified KVIs, which are strongly related to key value areas identified in [D1.2], are listed in Table **5**.

Table 5: 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)

Explainability and *fairness* are both essential components of trustworthy and ethical AI. The importance and reasons for requiring such features may vary depending on application areas. For example, in some cases decisions made by an AI system have direct impact on people's lives,

while in other cases the evaluation of technical feasibility and generalisability needs a suitable explanation of the outcome. During the design of these systems, often trade-offs must be made to provide better explainability or fairness, typically at the cost of accuracy. *Data economy* is another essential characteristic of ML systems due to multiple reasons. Obtaining the right amount and quality data for training can be challenging, especially when bias-free data with high diversity is required. Training on high volumes of data is also a computationally intensive process, which has significant impact on energy efficiency, communication resource efficiency and a sustainable operation, in general. Moreover, carefully selected training data may enable increased privacy, thus contributing to the trustworthiness key value area. *Model complexity* can be characterised in multiple ways. It can refer to the degrees of freedom of a learned model (e.g., number of trainable parameters), or the complexity of the model architecture (e.g., layer structure, memory properties), complexity of algorithmic learning, etc. These dimensions all impact practical performance measures, like accuracy, computational cost, training efficiency or model interpretability, as well as the relevant key value areas of sustainability by compute/energy efficiency and trustworthiness via explainability.

The use cases and KVIs identified above can be characterised (and quantified) by performance metrics defined by KPIs in Table 6. The targeted values of the mentioned KPIs will be proposed as part of future Hexa-X work.

Table **6**. 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

2.7 Challenges for data management, ownership, and privacy

Since generation, acquisition, transfer and processing of data are functionalities central to a future AI-enabled 6G network, this section aims to highlight some envisioned challenges regarding the management, quality, quantity, monetisation, privacy and integrity of data generated, exchanged and ingested by model-containing AI agents envisioned to be instantiated across a 6G network.

2.7.1 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. In other words, high quality data sets (e.g., leading to scarcer mis-labelling events in supervised learning) exhibit the advantage of accelerating model training convergence and, at the same time, enhancing the generalisation capability of the (updated) model. As it will be explained in detail in section 3.4.4 (and references therein), data points are characterised by different uncertainty levels, measured e.g., from the "distance" a given data point has from a decision boundary of e.g., a SVM [WLZ+19] [LZZ+19]. One may imagine that high-quality (or, "important") data points and data sets would be harder to seek and, therefore -assuming that data contributors would be provided feedback on the significance of their contributed data- their acquisition price would be higher. High quality data sets are also envisioned to positively impact the communication and computation resource efficiency of a network, while, not concurrently sacrificing the convergence speed of model training and the trained model's inferencing capability [ZLT+21].

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*); this measure is of importance to AI/ML-based network operation and overall performance of AI services, as the continuous transfer of generated data (network operation-related or application data) to deployed AI agents is expected to enhance the adaptability of the AI functionality to system/ environmental changes; along these lines [LXK+20] brings up a trade-off between age-of-data and AI service latency, that may be addressed differently, depending on the considered AI application. A characteristic example is the one of channel/ signal strength measurements: the more frequently these data points are generated, the more robust AI-based air interface functionality designs (e.g., channel estimation, beamforming) are expected to be to sudden radio channel condition fluctuations over time. On the contrary, slower data sampling is expected to relax data storage requirements and improve the energy efficiency of a transmitter/ receiver and a cache server, however, at the expense of lower system reconfigurability (or adaptability). Of course, there may be cases where "fresh" data may be unavailable for a period of time; to overcome this hardship that reduces AI agent availability (or, increases AI service "down time"), ML models may rely on already stored data or perturbations thereof, resulting to synthetic datasets. Such datasets may also be proven useful in the presence of new available training data, for bias minimization [LV19], as the "live" data may only be available from specific devices, locations or time intervals. As a bottom line, availability of timely data impacts how well a model represents the current (e.g., radio) environment. In relation to data freshness and regarding online learning, a crucial task consists in how to obtain a data set by proper time series sampling. The trade-off between data storage, needed communication signalling and sample processing, on one hand, versus model accuracy and timeliness needs to be studied.

Efficient (learning) data *lifecycle management* is also vital to efficient operation of an AI-enabled network. Data lifecycle management includes multiple functionalities, some of which are the

following: data curation [Sal+19], provenance (or, lineage, related to secure data traceability) [WSC21], data labelling by implementing self-supervised learning (useful for, e.g., air interface design, as in [Nad21]), active learning [CAL94], distributed storage per AI agents' needs and data deletion when data are deemed as outdated, e.g., per an expiration/ "staleness" criterion. Depending on the target application, suitable *metadata* may need to be included as data attributes to evaluate data freshness and indicate source origin and privacy level, information which may define a data point's eligibility for specific learning use. Availability of such metadata may enable automated in-network data curation tasks for learning, such as labelling, directing to the AI agent sharing same/ similar context and possibly others. Examples of such attributes are: time of data point generation and time interval of validity (possibly using intermediate validity steps, as e.g., critical applications may only use recently updated information, while less critical applications may also use older data, etc.), location of generation and others.

2.7.2 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.) [CVK17] [KBG+18] 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. It can be straightforwardly understood that 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. As an example, depending on the existence/ absence of a data management service acting as "middleware" between a data source and an AI agent (or, AI function), monetisation of learning data and their analytics can be direct or indirect. In another example, as models are based on processing combined multi-source data and it is envisioned to have multiple AI agents deployed over a network coverage area, a specific data set may be consumed multiple times by different AI agents containing different models - such use is expected to impact the monetary value of data, set by the data creator/ generator. In any case, reliability of the training data source needs to always be evaluated, as it could be the case that certain data generation entities could sell data sets of questionable quality, e.g., mislabelled data for classification purposes [GYY+20].

Three ethical and regulatory vectors involved in data monetisation, due to the conflicting interests of actors involved in the digital supply chain, are the following: (1) the *individual data creators* who generate files and records through their own efforts or own data-generating devices, such as sensors or mobile phones, have a claim to ownership of data; (2) the *business entities* generating data in the course of their operations, such as their transactions with financial institutions or risk factors discovered through feedback from customers also have a claim on data captured through their systems and platforms; (3) the *regulators* have started to address the issues related to data ownership for specific sectors of the economy.

Nevertheless, apart from data, ML models themselves can be monetised. In traditional approaches, the direct monetisation of data is considered, e.g., by transferring certain data sets from one party to another. In 6G systems and using the AIaaS principles, new approaches will arise for monetising data. In particular, a user may decide to request the creation of a “model” (such as the configuration of a NN following specific user requirements). In such a case, instead of training data being transferred to the model requestor, model parameters will be transferred by an entity containing a model (e.g., a FL aggregator). This approach, thus, leads to a different

learning vendor ecosystem, as compared to the existent ones. As this type of traffic might be increasing in future, standardised source coding and NN model compression becomes crucial to efficiently move models within the network.

2.7.3 Data privacy, security, and integrity

Data-centric AI/ ML-based solutions are expected to be extensively implemented within a 6G wireless communications system not only due to their promising positive impact in terms of performance, complexity and sustainability, but also harvesting the steadily increasing generation and exchange of data (both network and application-related) within the last years - data essential to fuel models for AI/ML-based air interface designs and distributed, federated and, overall, collaborative learning to serve future applications addressing 6G use cases. Nevertheless, 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 GDPR considering the "Ethics guidelines for trustworthy AI" [EC19]. To give an example on the GDPR compliance requirements, 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; for example, sensory and actuation data of a control system implemented on a factory setting involving robots and AGVs would need to be kept local and not depart a given deployment area. 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 [ABG+19]. 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. Therefore, 6G systems should support systematic features and services for data sharing, aggregation, de-personalisation and anonymisation. Collaborative and distributed learning methods are another enabling technology to leverage sensitive data minimisation in 6G systems [TSW+21].

With respect to the privacy of both learning and inferencing data, 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" and "untrusted" domains, which is the case with today's cellular networks. In 5G, the Network Exposure Function (NEF) indeed exposes capabilities and events. It stores the received information as structured data and exposes it to other NFs [23.501]. For this purpose, data will need to be stored and transferred applying a certain hierarchy structure, ranging from fully trusted to untrusted, including a number of intermediate levels of trust providing access to certain data sets upon authorisation. 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 [RB20].

Also, in the context of the newly activated Article 3(3)(e) of the Radio Equipment Directive (RED) on Privacy [RED], it is expected that further solutions will be mandated in order to protect stored, transmitted or otherwise processed personal data against accidental or unauthorised storage, processing, access, disclosure, unauthorised destruction, loss or alteration or lack of availability and also to implement appropriate authentication and access control mechanisms. Furthermore, under the newly activated Article 3(3)(e) of RED, it is expected that solutions must be implemented to address publicly known exploitable vulnerabilities as regards data protection and privacy, to include functionalities to inform the user of changes that may affect data protection

and privacy and to monitor and track the internal activity that can have an impact on data protection and privacy. Future 6G systems will be required to fully comply with the novel requirements.

2.8 Other related ICT-52 projects in the area

The ICT-52 5G-PPP program has funded several projects, including Hexa-X, within the framework Smart Connectivity beyond 5G. Thus, it is of interest to briefly present the objectives and methodologies of the most relevant projects with respect to the Hexa-X vision, to finally analyse common points, different views and scope of Hexa-X, and in particular WP4, in this area.

6G BRAINS

The goal of 6G BRAINS is to enable an AI-driven network architecture able to cope with highly dynamic environments and ultra-dense networks, as well as to devise new 3D indoor positioning methods. The methodologies involve AI, in particular bringing (Deep) Reinforcement Learning into wireless networks for cross-layer resource allocation purposes in massive machine-type communications, with new spectrum usage including THz and optical communications. In the 6G BRAINS view, the developed technologies will find applications to different vertical sectors, including Industry 4.0, intelligent transportation, and eHealth.

More details can be found at: https://6g-brains.eu/

AI@EDGE

AI@EDGE targets a closed-loop network automation for secure, reusable and trustworthy AI/ML models, and a connect-compute platform to support AI-enabled network applications. The methodology will build on different pillars, including AI/ML for closed loop automation, privacy preserving ML and decentralised connect-compute platform, AI-enabled applications, hardware-accelerated serverless platform for AI/ML, cross-layer, multi-connectivity and disaggregated radio access. Relevant use cases include vehicle cooperative perception, IoT networks secure and resilient orchestration, edge AI assisted monitoring of linear infrastructures using drones and smart content & data curation for in-flight entertainment services.

More details can be found at: https://aiatedge.eu/

DAEMON

DAEMON carries out a systematic analysis of which network intelligence tasks are appropriately solved with AI models, providing a solid set of guidelines for the use of ML in NFs. For those problems where AI is a suitable tool, DAEMON designs tailored AI models that respond to the specific needs of NFs, taking advantage of the most recent advances in ML. Building on these models, DAEMON designs an E2E network intelligence-native architecture for B5G that fully coordinates network intelligence-assisted functionalities. The main targets are to deliver extremely high performance, while making an efficient use of the underlying radio and computational resources, to reduce the energy footprint and to provide extremely high reliability beyond that of 5G systems.

More details can be found at: https://h2020daemon.eu/

DEDICAT 6G

Dynamic resourcing and connectivity to improve adaptability, performance and trustworthiness in the presence of emerging human-centric services are the main goals of DEDICAT 6G, which aims to develop a smart connectivity platform using AI and blockchain techniques to enable 6G networks to combine the existing communication infrastructure with novel distribution of intelligence at the edge. The proposal also targets the design and development of mechanisms for dynamic coverage extension through the exploitation of novel terminals and mobile client nodes, e.g., smart connected cars, robots and drones. DEDICAT 6G will focus on four use cases: smart warehousing, enhance experiences, public safety and smart highway.

More details can be found at: https://dedicat6g.eu/

MARSAL

MARSAL also falls under the umbrella of 5G and beyond intelligent networks, with several objectives, spanning from cell-free Massive MIMO (aligned with the O-RAN alliance architecture) to networking solutions such as Software Defined Networking (SDN) and cloud native solutions. Security, privacy, and trust are also main pillars of the project, as well as ML-driven control. The project is expected to produce Proofs of Concept (PoCs) in two areas, namely cell-free networking in dense and ultra-dense hotspot areas and cognitive assistance and its security and privacy implications in 5G and beyond.

More details can be found at: https://www.marsalproject.eu/

TERAFLOW

TERAFLOW aims at creating a cloud-native architecture, with SDN controller for beyond 5G network, integrated with Network Function Virtualisation (NFV) and Multi-access Edge Computing (MEC). Other objectives are to develop a ML-based security system and a trustworthy distributed ledger suing blockchain. The findings of the projects will be validated through final demonstrations supporting three different network scenarios, namely autonomous networks beyond 5G, automotive and cybersecurity.

More details can be found at: https://www.teraflow-h2020.eu/

Differences and similarities with Hexa-X

Summarising the ICT-52 expected contributions in the area, we can identify a common agreement and convergence on trustworthiness and security aspects, as well as a joint view of communication, computation and AI/ML aspects into wireless networks, as also highlighted in Chapter 4 of this document. AI-based architectures, which are the main pillars promoted by WP4, are strongly encouraged by the majority of the activities of other related projects. Sustainability is, however, one of the central topics of Hexa-X that is less stressed in other projects. Hexa-X, targeting a flagship vision, aims to jointly address the challenges of trustworthy AI /ML, heterogeneous resource aggregation, sustainability, coverage and extreme experience, rather than focusing only on specific aspects of AI (e.g., trustworthiness), which is the main focus of WP4 work.

2.9 Overview of technical enablers for AI/ML in networking

Following the elaborations in the previous sections, Figure 2-4 below illustrates the main research areas covered in this deliverable, namely, *AI-driven air interface design* and *In-network learning methods and algorithms*. Within the first research area, the following more specific technical enablers are elaborated: (a) novel, data-driven transceiver design approaches, (b) AI-driven transmitters, (c) AI-driven receivers and (d) AI-driven radio interface functionality. On the other hand, the second research area includes technical enablers grouped as follows: (a) joint communication and computation co-design as enabler for distributed learning, (b) enablers for in-network AI security, privacy and trust, and (c) AI-powered network operation. On top of Figure 2-4 the initially selected (sourced from [D1.2]) or new suggested KPIs and KVIs are mentioned, relevant to each of the two research areas.

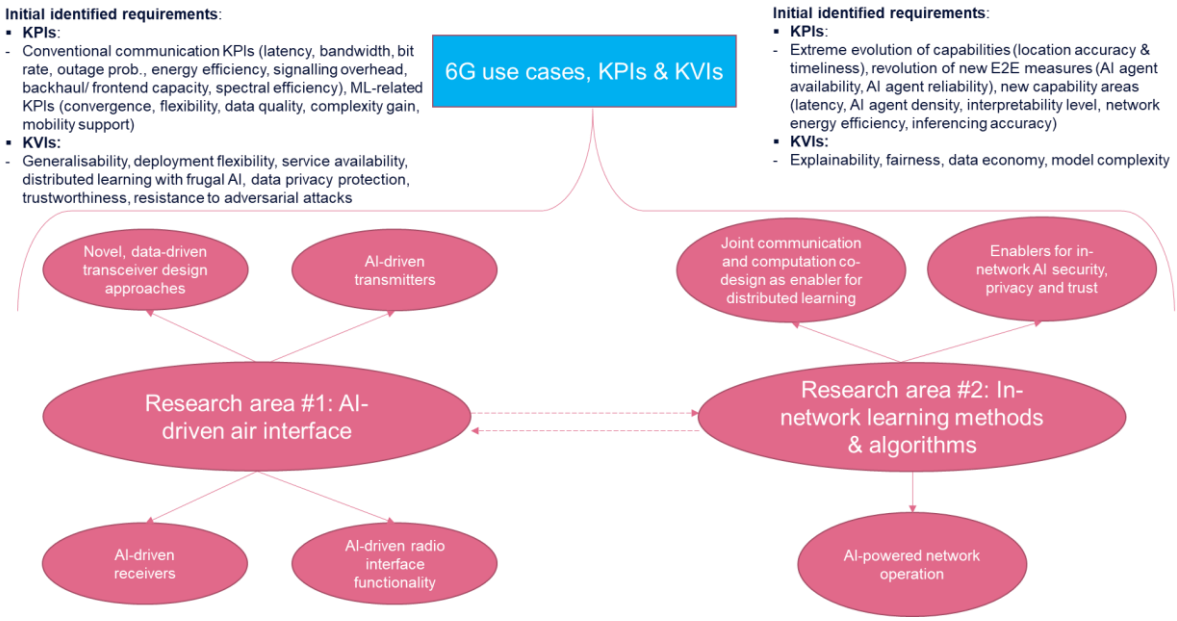

Figure 2-4: Overview of technical enablers

In what follows, Chapter 3 covers the technical enablers within the *AI-driven air interface design* research area, while Chapter 4 focuses on the technical enablers within the *In-network learning methods and algorithms* research area.

3 Motivation and gaps of an AI-driven air interface

This chapter focuses on AI-based air interface design via four themes. They are as follows: novel, data-driven transceiver design approaches, AI-driven transmitters, AI-driven receivers, and AI-driven radio interface functionality. These fall broadly on the physical (PHY) layer techniques and radio resource management (RRM) with AI and ML. Hardware impairments in the transmitter and receiver RF chains play a significant role in the design as to how to take those into account. This is important given that we consider the lower mmWave (30-100 GHz) and upper mmWave (100-300 GHz) frequency ranges and not many hardware impairment models are widely available. Associated with the transmitter and receiver is also the MIMO antennas and hence their beamforming is considered via AI optimization as well as fast initial access with RL methods. These are evaluated for multiuser systems. In the receiver side, various methods of channel estimation schemes are studied for lower complexity including those in RIS systems. Careful analysis of latent variables (i.e., variables not directly observable, but rather inferred from other variables that are observed) in channel estimation is important as they impact on RRM algorithms. As an integral component in the receiver side, channel decoding is explored with ML algorithms. Receiver side processing as a single block is also investigated considering hardware impairments as well. Furthermore, RRM is a crucial aspect where different methodologies are proposed. Cell-free and distributed massive MIMO systems are under investigation. Along with this data importance aware RRM where data quality can be quantified in several measures need proper analysis. Finally, model predictive control of antenna systems is considered with batch and online ML and RL algorithms. Figure 3-1 illustrates an overview of the proposed enablers for AI-driven air interface design and achieving the proposed 6G use cases via them. Figure 3-2 illustrates the focused 6G use cases, as selected from [D1.2] and elaborated in section 2.6.1 and clarifies upon their importance and challenges from an AI-driven air interface design standpoint, whereas, Figure 3-3 provides, on top of section 2.6.1, an overview of identified KPIs and KVIs related to the selected 6G use cases, that are of utmost importance to evaluate the efficiency of AI-based air interface designs.

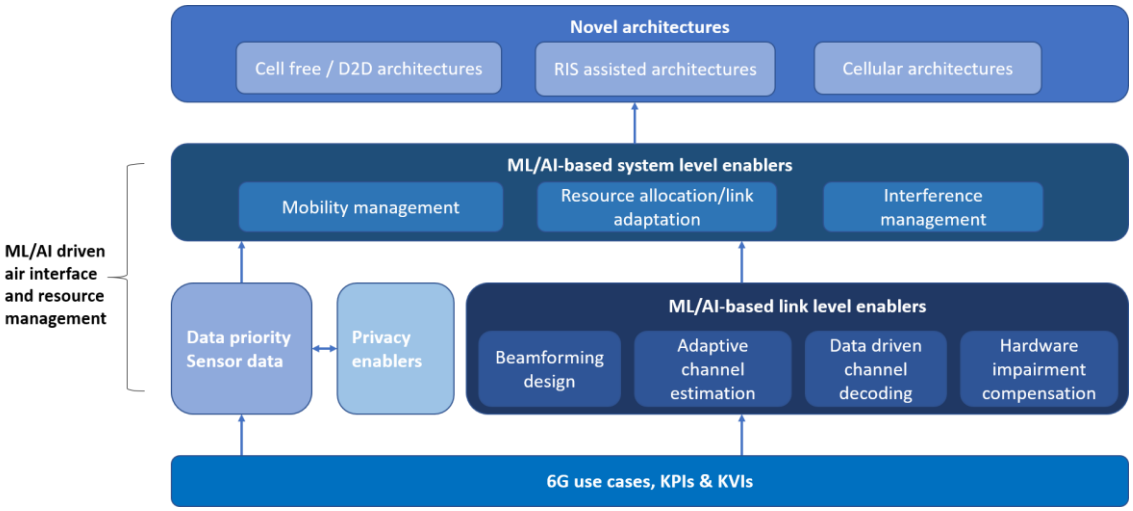

Figure 3-1: Overview of proposed enablers for AI-driven air interface design and connection to the 6G use cases

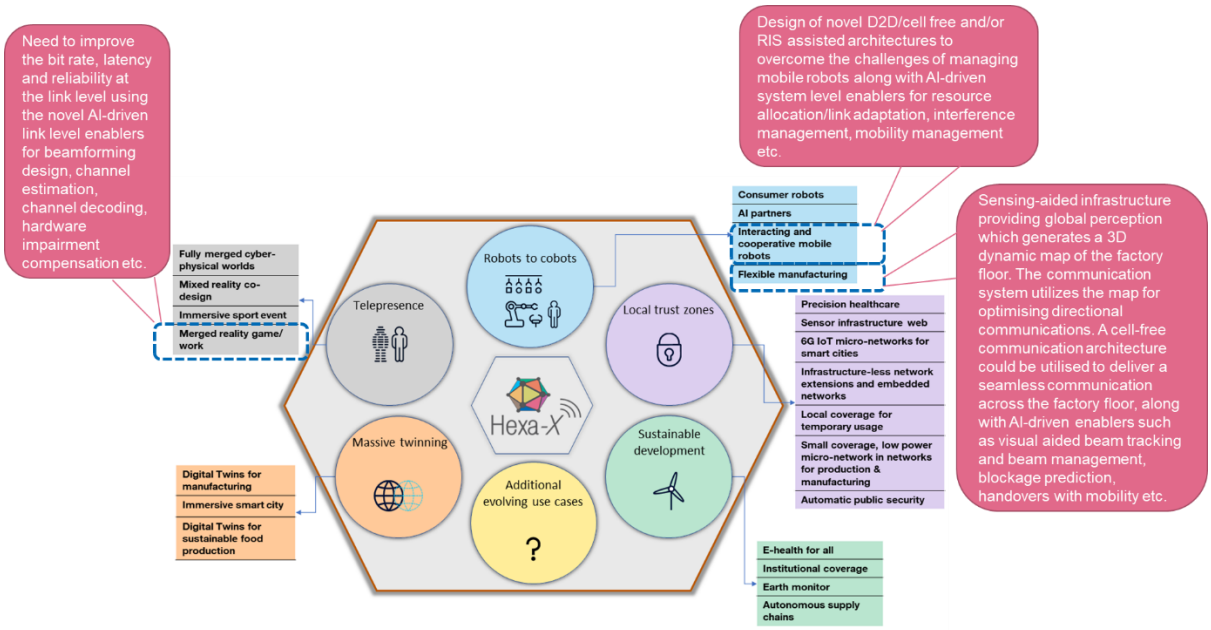

Figure 3-2: Tailoring technology enablers on AI-driven air interface to identified Hexa-X use cases, based on [D1.2]

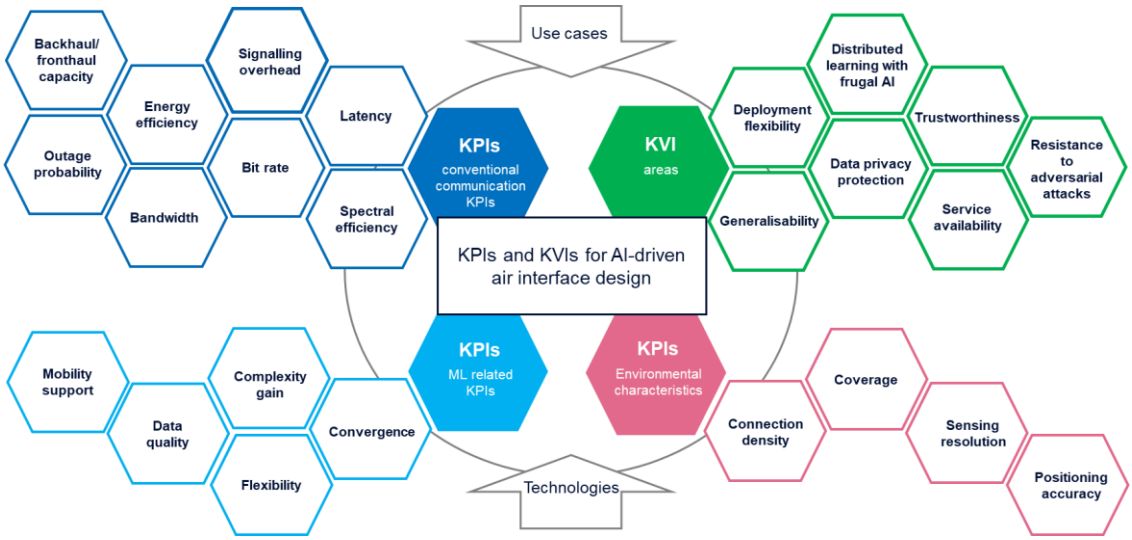

Figure 3-3: Overview of identified KPIs and KVIs for AI-driven air interface design

3.1 Novel, data-driven transceiver design approaches

3.1.1 Transceiver hardware impairments

6G is envisioned to enable transmission at higher carrier frequencies and to provide opportunities with larger spectrum to be allocated for transmissions [BC20]. However, transmission at these frequencies compared to the already existing frequency bands would be even more challenging due to the prominent impact of hardware impairments. Transmitted and received signals are subject to distortions due to different sources of Radio Frequency (RF) hardware impairments in the transmitter's and receiver's RF chains such as oscillator's phase noise or power amplifier nonlinearities on link performance. These distortions would lead to the deviation of the transmit signals from the intended designed signals and, hence, contribute to performance degradation.

Conventional approaches have focused on mitigating RF hardware impairments by designing more sophisticated RF hardware components or try to design algorithms to compensate specific type of RF hardware impairments. However, this approach is not scalable and economically viable, as RF circuit design for operation at higher frequencies becomes even more challenging and costly. Hence, the communication systems would be required to operate in the presence of more severe distortions due to RF hardware impairments. The operation of the conventional algorithms in the presence of RF hardware impairments becomes sub-optimal indicating that there is potential for system performance improvement by designing AI/ML-based methods that can take into account hardware impairments during the communication system operation.

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) optimizing the operation of receiver and/or transmitter algorithms based on the inferred information about the hardware impairments. An example solution, based on this approach has been developed for phase noise compensation at the receiver for high band transmission in [FS20] leading to performance improvement compared to the existing methods.

Methods to be developed for handling hardware impairments with AI/ML in radio transceivers will help to address communication KPI requirements, especially in the high frequency regime, where device specific impairments are expected to significantly influence the achievable performance. Without such mitigation methods and without the use of AI/ML techniques fulfilling the conventional communication KPIs in the high frequency range may not even be possible.

3.2 AI-driven transmitters

3.2.1 Beamforming design, beam management, multi-antenna signal transmission

Higher frequencies such as mmWave inherently suffer from high propagation loss and scattering. Hence innovative beamforming designs and efficient beam management procedures with multi-antennas need to be developed. Here beam management procedures refer to initial attach, beam recovery and beam steering. 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. Beam management has to be designed such that it causes the least amount of overhead in terms of time since mmWave channels have short coherence times. Very narrow beams will be used in higher frequencies like mmWave and communications will mostly resort to having the line-of-sight path. Given the above discussion, beam alignment methods have to be efficient and fast to keep up with the changing circumstances. Therefore, novel innovative methods are required to handle beamforming design and beam management efficiently.

Recently, researchers have focused on using ML based techniques to assuage the complexity involved with future communication systems [ASR20, IMA21, IAR21, AA20, ZAA21]. Authors of [IMA21] introduces a Deep Reinforcement Learning (DRL) based novel approach to provide fast beam search for initial access (IA) in cell-free dense mmWave networks. A single user reference pilot is used to predict the downlink beam using the DRL model. Synchronization part has not been considered in [IMA21]. The work in [IAR21] presents a deep contextual bandit (DCB) based approach for access points to provide initial access to users using beam management information from their neighbouring access points. For heterogeneous deployment scenarios,

authors of [AA20] proposed using sub-6 GHz channel information to predict beams and blockages for mmWave channels. However, solution in [AA20] cannot cater standalone deployment of mmWave networks. Authors of [ZAA21] suggest a novel DRL based method to customize the beam codebook design to a given environment. Each beam is considered as an agent and the number of agents corresponds to the predefined number of beams. Optimization of the number of beams is interesting. Cell free and its extension to user centric architectures have been proposed as an alternative to the conventional cellular system to meet the high throughput and reliability demands of the future communications networks [NAY17,BA17]. Here users are simultaneously served by multiple access points (APs) which are controlled by a central controller.

Although communication range is an issue in mmWave, cell free architecture can help mitigate this with dense AP deployment scenarios. Cooperative beam management for the uplink and downlink in the mmWave user centric networks is complicated. Initial access concerning control plane must provide beam identification, resource allocation and user association at the same time. With regards to fast and efficient beam management pertaining to data plane in cell free architecture, novel ML based hybrid, or low-resolution digital beamforming algorithms can be developed for cooperative serving of the user.

Furthermore, sensing the environment through other modalities than RF emerge as a complementary technology to capture the dynamics which can imply information on the scattering thus aiding in the communication procedures such as beam design, beam management and blockage prediction. Regarding sensing, the work in [JRL19] introduces an infrastructure mounted LiDAR based system for communications in vehicular scenarios which can be applied in indoor scenarios as well. It also demonstrated a practical method for creating a 3D map of the environment using such a system and navigation of a robot through the 3D map with minimal sensing as a proof of concept [PMI21]. Cameras, LiDARs and radars are such sensors that can be used. When LiDARs are considered, in [KGH19] and [DKG19], a deep learning based mmWave beam selection problem has been explored which utilizes LiDAR data in vehicles. Use of camera data has been explored for mmWave beam selection in [XGJ20] and [CAA20]. Another problem considered in literature is the dynamic blockage prediction which can aid in beamforming and beam management in mmWave networks. In [CAA20], camera images obtained through cameras fixed on the base station are utilized to predict whether the user will be blocked. However, these studies related to sensing though other modalities still have preliminary work which should be studied further for different scenarios and specially with mmWave systems.

As discussed previously, mmWave communications inherently suffer from susceptibility to random blockages which can degrade the link quality drastically. This is due to the use of narrow beams in mmWave communications and the line of sight in these beams getting blocked. The random blockages can be self-blockage or other entities like dynamic movement of people. Predicting the blockage and allowing enough time for the transceiver to act proactively can be enabled by sensing the environment. Furthermore, sensing can aid in beam design, beam searching and tracking. Therefore, exploration of utilization of sensing data from LiDARs and radars coupled with ML/AI-based methods for beam design and beam management will be performed.

The mostly relevant use case for this contribution would be the flexible manufacturing use case where additional infrastructure-based sensors such as LiDARs would be used to assist the navigation and communications in the factory floor. Cell free architectures with AI-driven visual aided beam tracking and beam management, blockage prediction and handovers with mobility are some of the enabling technologies in this regard. In addition to the conventional communication KPIs and ML-related KPIs, environmental characteristics such as the coverage and sensing resolution are also to be considered.

3.2.2 AI for multi-cell, multi-user MIMO

As cellular data demand continues to rise, ultra-dense networks are widely considered as a key component in managing this trend. MIMO solutions have been developed for efficient transmission and reception of radio signals in multi-antenna systems. In particular, downlink multi-user MIMO is a promising technique to achieve higher throughput in a multi-cell environment. However, in general, the optimization 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.

Reinforcement learning is a set of ML techniques that allows an agent to learn the optimal action policy that returns the maximum reward through trial-and-error interactions with a challenging dynamic environment [SB17]. RL has been used to solve challenging problems in various areas ranging from games to robotics. In [LGJ1-20] [LGJ2-20] different RL-based approaches are investigated to improve the performance of MIMO systems. However, these studies have focused on enhancing the performance of single-cell MIMO systems.

Multi-cell, multi-user precoding problems can be seen as a multi-agent system that learns to coordinate transmission schemes (or action policies) in interaction with other base stations (or other agents). The multi-agent problem requires complex inter-cell interference coordination in the sense that each BS should exhibit cooperative behaviour to maximize the signal power to a desired user while minimizing the interference power to other users in the multi-cell environment. This problem poses two main challenges: i) multiple actors (or agents) with partial observability and ii) multi-dimensional continuous action space. The first challenge is a direct result of practical limitations of accessible information by local agents distributed in MIMO interference channel, and the second challenge comes from the fact that multi-dimensional precoding vectors should be optimized for multi-antenna BSs based on a certain transmit power constraint.

Therefore, scaling and adapting previous works on multi-agent systems to real-world, multi-cell, massive MIMO environments is crucial to building future intelligent networks. The application of multi-agent actor-critic models [LWP17] to achieve cooperative transmission from multiple base stations can be one approach to learn how to solve the complex multi-point transmission problem via trials and examples. To address the two main challenges: multiple actors with partial observability and multi-dimensional continuous space in real-world cellular systems, we adopt a multi-agent deep RL (MADRL) framework, as illustrated in Figure 3-4 where decentralized actors with partial observability can learn a multi-dimensional continuous policy in a centralized manner with the aid of shared critic with global information. The framework allows for centralized learning with decentralized execution at different levels of observability and time requirement, which is a realistic and practical approach for real-world cellular environments.

AI/ML techniques developed for multi-cell, massive MIMO environments would enable to utilize the capacity potentials of the massive number of antennas and fulfil the conventional communication KPIs. For example, regarding the "Flexible manufacturing" use case, massive multi-antenna systems are expected to be used to provide the high reliability and capacity requirements. Avoiding interference and utilizing the spatial channel capacities with ML/AI optimized multi-antenna transmissions will be of high relevance in this scenario.

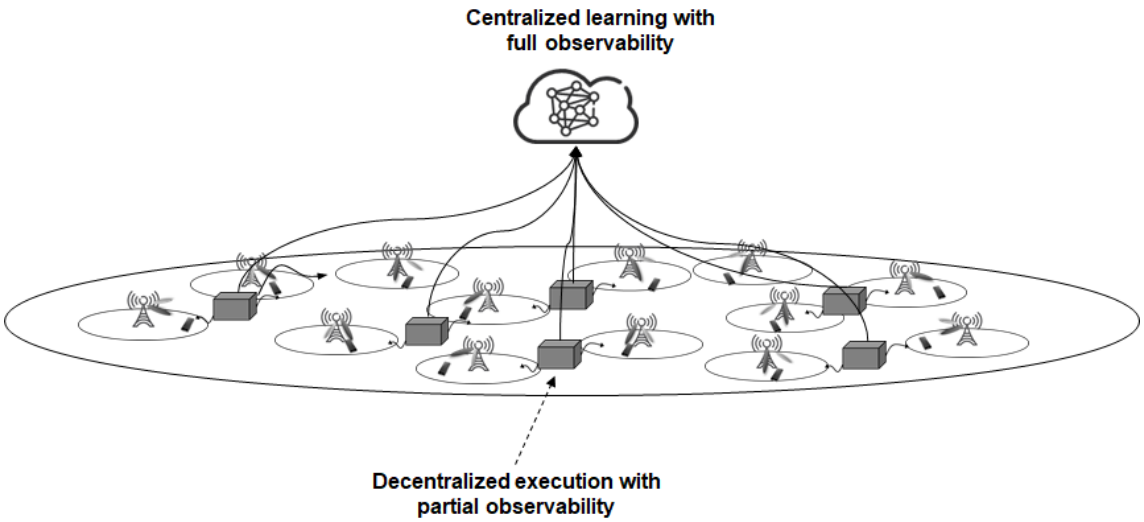

Figure 3-4: Multi-agent Deep RL framework for centralized learning with decentralized executions

3.3 AI-driven receivers

3.3.1 Adaptive channel estimation/denoising

The problem of interest here is the mitigation of hardware impairments for MIMO channel estimation based on a physical model. Indeed, channel models are nothing but imperfect representations of the channel manifold relying on simplifying assumptions. Some popular assumptions are for example the isotropic antenna model and the plane wave assumption. In order to optimize channel estimation using physical models, it is required to handle uncertainties about the antenna gains, positions or radiation patterns (among other system features).

Dictionary learning is a method allowing to adapt a sparsifying basis to incoming data, and is thus suited to the channel estimation task in which the channel is often assumed to be sparse in a basis of steering vectors. Classical dictionary learning has indeed been applied to channel estimation [DR18] in order to mitigate such impairments, but in an offline and computationally expensive fashion. It has been later proposed to use an online dictionary learning [MBP+10] approach based on a deep unfolded neural network [HRW14] to tackle the problem. The resulting method called mpNet [MP20] allows to correct online an imperfect model (adapting to the actual channel manifold) and is totally unsupervised, being trained as an autoencoder. In a follow-up work [YL20], it was also proposed to automatically adapt to the signal to noise ratio (SNR) by dynamically changing the network's depth. Such a method is also shown to be resilient to incidents at the base station (such as antennas being displaced or broken by accident), because it allows to adapt in real time to the radio environment. An example of channel estimation performance obtained over time with mpNet at an SNR of 10 dB is shown on figure 3-5, where it is compared to several baselines, and an incident occurs at a specific time during the simulation. The blue curve corresponds to mpNet. The green one corresponds to a physical model for which uncertainties are not taken into account. The red one corresponds to use simply least squares estimation, and lastly the orange one corresponds to a physical model that is perfectly calibrated initially. The main message is that the resilience brought to mpNet by online learning makes it adapt over time in order to optimise performance, even if the system features are suddenly affected by some incident, which is not the case of concurrent methods.

The main objective is to pursue previous work with the aim to handle more naturally the channel manifold in the future (without the implicit discretization involved in dictionary learning), in order

to improve the accuracy/complexity trade-off. This will be done, for example, using advanced representation learning methods (such as variational autoencoders [KW13]). Compared to previous work, this would imply changing the neural network structure, without changing the training objective.

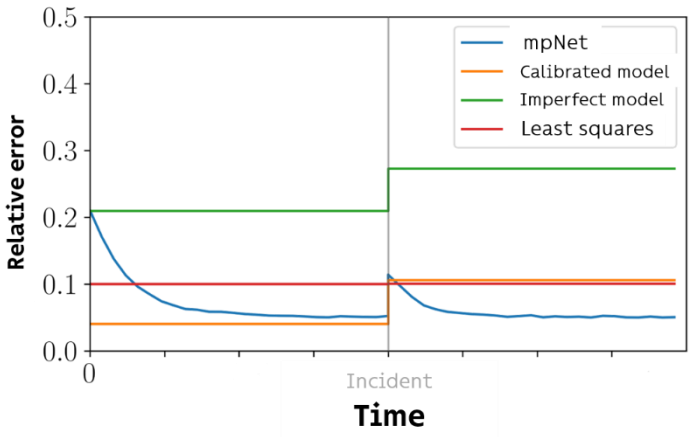

Figure 3-5: Example of obtained result over time with mpNet and several baselines

3.3.2 ML-based channel estimation for RIS assisted systems

Reconfigurable intelligent systems (RIS) is an emerging technology in modern communication systems to improve the performance in propagation environments with severe blockage or in scenarios where propagation range is a challenge like mmWave communications. Channel estimation of RIS–assisted systems is a research area that needs to be addressed since an RIS contains a large number of elements, it increases the number of links to be estimated and 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.

In the existing literature, there are two main approaches for RIS channel estimation based on two different RIS configurations, depending on whether the RIS is mounted with sensing devices (receive RF chains) or not, termed as semi-passive RIS and (fully) passive RIS, respectively. In the semi-passive scenario, there are sensors in RIS to help estimate the channel, but it is not a cost-efficient method as there are many sensors needed. When there are no sensors mounted on the RIS, RIS becomes fully passive and thus it is generally infeasible to acquire the CSI between the RIS and BS/users directly. Therefore, only approach is to estimate the cascaded user-RIS-BS channels at the BS/users in the uplink/downlink, respectively [WZZ+21]. One practical method for RIS channel estimation is by employing an ON/OFF-based RIS reflection pattern i.e., each one of the RIS elements is turned ON sequentially with the others set OFF at each time, thereby the user-BS direct channel and the cascaded channels associated with different RIS elements are estimated separately [MJ19]. Authors in [JC20] show that this ON/OFF method is suboptimal because in the optimal scenario all elements should be in ON state and by considering the full reflection of the RIS during full channel estimation phase it can achieve optimal channel estimations if the RIS reflection patterns follow the rows of a discrete Fourier transform (DFT). The authors in [ZZ19] also used the above method to estimate the channel of RIS assisted orthogonal frequency-division multiplexing (OFDM) system. They have shown that it is needed at least Number of elements in RIS + 1 pilot symbols to estimate the channel. However, expected channel estimation accuracy cannot be met with conventional channel estimation techniques. Hence ML/AI-based techniques can be used to have more efficient channel estimation algorithms

for RIS -assisted systems. The authors in [KM21], consider the design of optimal channel estimation strategies for RIS-aided multiple antenna communications, based on the minimum mean squared error (MMSE) criterion with the help of data-driven deep learning approaches.

Unsupervised deep learning without the need for offline training (there is no need to use a large dataset with labels) approaches can be used to come up with efficient algorithms for channel estimation for RIS-assisted OFDM systems and reflection pattern optimization. It can be extended to explore scenarios, such as systems with high mobility (vehicular) users and/or high mobility RIS, systems which have multiple RISs and more realistic systems with hardware impairments. Thus, ML/AI-based channel estimation would be an enabler for the interacting and cooperative mobile robots use case where RIS assisted systems are used to manage a cluster of drones over a 6G network. In addition to achieving the traditional communication KPIs, ML-related KPIs such as complexity gain, flexibility, convergence and mobility support will be considered when determining the performance of the proposed algorithms.

3.3.3 Low complexity channel estimation

Precise channel state information (CSI) for transmit precoding and beamforming is crucial. The problem of acquiring accurate channel data with minimal signalling overhead becomes extremely relevant for massive MIMO (mMIMO) scenarios. 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). Even then, the problem of sample complexity and computational complexity is prevalent. Due to the non-linear nature of the problem, neural networks (NNs) might be interesting candidates to investigate.

A NN-based channel estimation approach is proposed in [NWU18], in which the conventional MMSE channel estimator is equivalently represented under certain array-specific property assumptions as a two-layer NN. The conventional channel estimators are usually more complex than the ML-based solutions, e.g. [DZL+19], even though they achieve similar performance. For mMIMO systems, the computational requirements will be further increased [MAR10]. Another proof-of-concept is provided in [HDW+19], in which a ML-based solution is tested on measured data. In [KHU19], on the other hand, ML-based channel estimation is considered for mMIMO with different antenna configurations.

For any future practical channel estimation algorithm, three aspects are of high importance: sample complexity (the number of samples required to learn and generalize well), computational complexity (in terms of flops per inference), and accuracy. However, treating this problem only on the statistical and/or machine learning level might pose difficulties in practice in terms of explainability. Therefore, methods should be proposed that can benefit from the domain expert knowledge to guide the optimization procedure for the NN in the right direction. Domain knowledge should be also put in use to produce systems and procedures for a low-overhead online fine-tuning of the solution. A study on comparing training schemes such as pure online learning, offline learning with online fine-tuning, and pure offline learning is extremely necessary for applications such as channel estimation, where all the three methods could be applied. In addition to the aspects mentioned above, one must consider the following two additional aspects: computational complexity of online training and signalling overhead required to collect data and labels for online training. With future networks envisaged to be even more heterogeneous than 5G, device type and application is crucial in choosing the correct scheme.

3.3.4 Data-driven channel (de)coding for constrained devices

6G is envisioned to address the expansion of Internet of Things (IoT) use-cases whose traffic and QoS requirements are deemed to grow significantly over the next decades. Efficient and low

complexity codes and associated decoders for intermediate length datagrams - typically a few hundred bits - are key requirements to address this trend without exhausting the available spectrum resources nor surpassing the capabilities of IoT constrained devices. To this end, Neural Network models and Machine Learning optimization techniques are promising tools to design, optimize and decode codes that fall in between long codes, for which efficient schemes are known, e.g. Low Density Parity Check (LDPC) codes, and robust short codes, e.g. 5G Polar Codes.

In the state-of-the-art, a particularly interesting "clear-box" functional approach (see section 2.3.2) was recently proposed by Nachmani et al. [NML18]. The proposed Neural Belief Propagation (NBP) decoder introduces a promising and explainable way to improve the decoding performance of a belief propagation iterative algorithm for short to medium length linear block codes for which the high density of the parity check (PC) matrix and presence of short cycles usually degrade the performance of the decoder. This new decoder represents the graph structure of BP algorithm as a NN, therefore enabling adaptive weighting of the messages exchanged during the decoding process. The authors showed in simulations that this NBP approach improves the decoding performance for short to medium block-length linear bock codes such as BCH codes. Several publications discussed improvements of NBP using pruning [BHP+20a], [BHP+20b], weights sharing [LCH+19] or active sampling [BRR+20]. NBP algorithm was also proposed to improve performance of permutation or list decoders where several BP decoders are executed in parallel with different permutations and showed to approach maximum likelihood decoding performance although at an additional cost in terms of complexity [NML18]. Yet, to the best of our knowledge, all state of the art uses of NBP algorithm assume a perfect knowledge of an expert-designed coding scheme (not necessarily optimal) on which rigorously depends the NN architecture.

The state-of-the art also include several publications adopting a "black-box" functional approach with complex, deep neural network models at the expense of a higher number of parameters and a reduced explainability. As an example, Deep Neural Network decoders are proposed to learn the decoding of convolutional and turbo codes [KJR+18] and polar codes [CGH+17]. Some interesting deep learning approach push this idea further, leading to hybrid LDPC/polar decoders [WZZ+18].

Contributions of the state-of-the-art exhibit different approaches regarding the question of the training of coding and decoding models: gradient descent of a loss function, and reinforcement-based methods. The first category, which encompasses most of the contributions can further be categorized by the figure of merit represented by the loss function: either a systemic - or end-to-end – performance criterion such as bit error rate, block error rate or similar metrics (mean square error, binary cross entropy, etc.), e.g. [NML18], or a derivative figure of merit of a part of the system, e.g., the mutual information between received samples and coded messages [FSW19]. The second category contains publications trying to circumvent the issue of learning with non-differentiable models such as digital modulations (BPSK, QAM for example) or arbitrary channel models [AH18].

NN and ML are often presented as a way to improve the adaptability of the air interface. Data driven approaches could further improve decoding processes and, notably, NBP decoders, by allowing the learning of the decoding process without prior knowledge of the coding scheme used and thus enabling blind NBP decoding of linear block codes. This raises, among others, the question of the learning of the topology of BP decoding graphs. The use of conditional computation mechanisms such as the one present in gated architectures (Long Short Term Memory [HS97], Gated Recurrent Unit [CMG+14]) could be of great interest to perform such task. The use of structured graph architectures inspired by the standard BP algorithm could lead to parsimonious models of reduced complexity, improved explainability, scalability and possibly equipped with better generalization capabilities for larger codes [GCH+17], [GBC16] where

exhaustive dataset cannot be used. On the other hand, over-structured architectures could impact trainability, expressivity and potential performance gain. Permutation decoders could also benefit of NN and ML. Such decoder could learn to choose a single relevant permutation depending on the received code-word, to bring the performance of a permutation decoder at the cost of a simple BP or NBP decoder.

Nonetheless, the performance of a communication chain not only lies in the decoding process but also in the choice of a performing coding scheme. One step further in the application of a data driven approach to the question of decoding would be the co-design of both coder and decoder which might partially be enabled by the aforementioned adaptive decoding approach used in an end-to-end model as an in-place replacement for the corresponding coding and decoding parts. State of the art currently exposes only a few publications concerning this topic which could be of great interest [HZL+19]. Once again, this approach could be extended to the question of permutation decoders with simultaneous design of the coding and decoding scheme including the design of relevant permutation and associated permutation selection mechanism.

In a fully explainable and model-based approach, the translation of encoding and decoding algorithm as dedicated NN models would also require extensive use of finite fields (notably binary finite field) related operations such as Binary-XOR. In a gradient descent-based training, one should account for ML related constraints such as differentiability of these operations. Reinforcement learning methods could be another interesting approach to bypass these problems and learn in non-differentiable parameter spaces.

AI and ML technics applied to channel (de)coding might lead to the improvement of conventional communication KPIs such as BER, latency and energy efficiency. These techniques would also benefit of taking into account AI/ML specific KVIs such as explainability, model complexity as well as inference and training complexity.

3.3.5 Toward an end-to-end driven receiver design

When applying ML techniques and architectures to physical layer radio reception, one should carefully consider the balance between expert knowledge and freedom of the ML algorithm to learn the best, and possibly unforeseen, solution. In certain cases, utilizing expert knowledge will likely lead to the most efficient solution. For instance, if one limits the ML algorithm to learn a very restricted task, it might be unlikely that a performance gain over conventional expert knowledge-based systems will be observed. In such a case, the only benefit one can expect to obtain is a reduction in computational complexity. While it is possible to avoid this by going to the other extreme and train the ML receiver to carry out several tasks without any restrictions, this might also be suboptimal since there are processing stages for which the analytically optimal method is known. A key aspect is, therefore, identifying the proper balance between prior knowledge/restricted tasks and the freedom to learn several tasks jointly. In particular, one should consider the complete end-to-end radio link, from the transmitter to the receiver, and identify which parts of the system should be learned from data. Such end-to-end driven receiver design scheme will identify the receiver architecture candidates that are suitable for a native ML air interface, where also some aspects of the transmitter could be learned from data.

ML-aided radio receiver design and operation has already been studied in several works, many of which have investigated implementing certain parts of the receiver chain with trainable layers or NNs. For instance, NN-based channel estimation has been studied in [NWU18, CMW+21, HWJ+18], while [CWL+19] proposes an equalizer consisting of a convolutional neural network (CNN) [LBH15]. ML-based demapping has been studied in [SH19], where it was shown to achieve nearly the same accuracy as the optimal demapping rule, albeit with greatly reduced computational cost. There are also some works that propose augmenting the conventional receiver

processing flow with deep learning components [GJW+18, HJW+19, SDW19] and show improved performance in comparison to fully conventional benchmark receivers.

In addition, completely ML-based receivers have also been proposed in the prior art. Especially, in [HKH21, KHH+21], a fully convolutional NN-based receiver, DeepRx, was proposed. It was shown to achieve high performance, especially under sparse pilot configurations. Learning larger portion of the receiver is also investigated in [YLJ18], where channel estimation and signal detection is carried out jointly using a fully connected NN. The proposed ML-based receiver is shown to outperform the conventional receiver when there are few channel estimation pilots or when there is no cyclic prefix being used. It is also shown to be capable of dealing rather well with clipping noise, a type of hard nonlinearity. The work in [ZVG+18], on the other hand, implements a CNN-based receiver which extracts the bit estimates directly from a linear time-domain RX signal.

The prospect of learning the transmitter and receiver jointly has also been investigated in the prior art [OH17, DCH+18, AH21, FCD+18]. Such schemes do not assume any prespecified modulation scheme or waveform, but instead learn everything from scratch. Such end-to-end learning has been shown be able to outperform traditional heuristic radio links, e.g., by learning a better constellation shape [AH21] or by learning to communicate under a nonlinear PA [FCD+18]. However, such a scheme must also learn such components or aspects of the radio system whose optimal solution is known a priori, therefore making the overall learning task less efficient.

Despite the wide body of literature regarding ML-based radio receivers (and transmitters) and the various demonstrations of their high performance, there are still various research gaps that should be addressed. Firstly, the proper trade-off between expert knowledge and freedom of learning should still be further analysed, i.e., which of the tasks are beneficial to be learned jointly, and which should be given as *a priori* knowledge. Secondly, the effects of different hardware impairments have been largely omitted in the analysis thus far. This aspect requires further research, as it is possible that fully or partially ML-based transmitters and receivers facilitate more efficient communication in a different operation point than the current systems, e.g., in terms of transmitter linearity requirements. Lastly, the computational complexity, in terms of bps per Watt per euros, should also be considered in the analysis, to determine which scheme is the optimal one for the operator.

The performance of ML-based radio links, which include learned components in transmitter and/or receiver, should be primarily measured based on the conventional communication KPIs, such as the achieved bit rate, spectral efficiency, energy efficiency, signalling overhead and latency. However, the ML-related KPIs have also some significance, especially to the vendors and operators, who are the ones training and deploying the models. Faster convergence, lower complexity, generalisability and more flexible deployment are crucial KPIs and KVIs from the business perspective. As for the UCs, the designed physical layer ML architectures should be generic enough to support heterogeneous deployment environments.

3.4 AI-driven radio interface functionality

3.4.1 Radio resource management based on channel latent variables

The problem of interest here is pretty general, and aims at easing RRM using latent variables learned from measured channels. Such latent variables can be for example angles of arrival and path gains, any estimated parameters from a physical model or activations of some hidden layer in a deep neural network (such as an autoencoder for instance). The applications potentially concerned by such an approach are countless, but the first envisioned concrete applications are:

(i) user positioning, (ii) channel mapping (suppressing the need for downlink pilots in frequency division duplex (FDD) systems), (iii) channel charting (unsupervised relative positioning of users, easing resource allocation or handover management for example).

Channel mapping has originally been tackled with a generic neural network structure [AA19]. This leads to a computationally heavy solution requiring a huge amount of training samples. Using instead a specific structure (such as the similarity-based neural network introduced in [LeM21a]) leads to a better performance/complexity trade-off. Moreover, it allows to handle both channel mapping and user positioning in a unified way. Channel charting was originally proposed based on high-dimensional features (square of the channel dimension) [SMG+18], which hinders its applicability. A more computationally efficient method based on a specifically designed distance measure [LeM21b] seems promising, see figure 3-6 for an example of obtained chart in an urban environment. Dots being grouped by colour on the chart means that charting is successful in preserving the local geometry of the channel manifold.

Based on previous work, the objectives of this contribution are: (i) to further evidence the interest of such methods in concrete applications (for example by designing experiments in more realistic conditions) and (ii) to design an unified approach with previous contributions on channel estimation/denoising [MP20]. This would lead to learn a global model of the channel manifold, in the sense that mappings from the observation space (channel measurements) to the latent space (coordinates on the manifold) and from the latent space to the observation space would be learned jointly. More concretely, once the manifold is properly modelled with encoder and decoder functions, projecting on the manifold (full pass in an autoencoder network) would correspond to channel estimation/denoising and computing latent variables (pass in the encoder only) could be used for RRM.

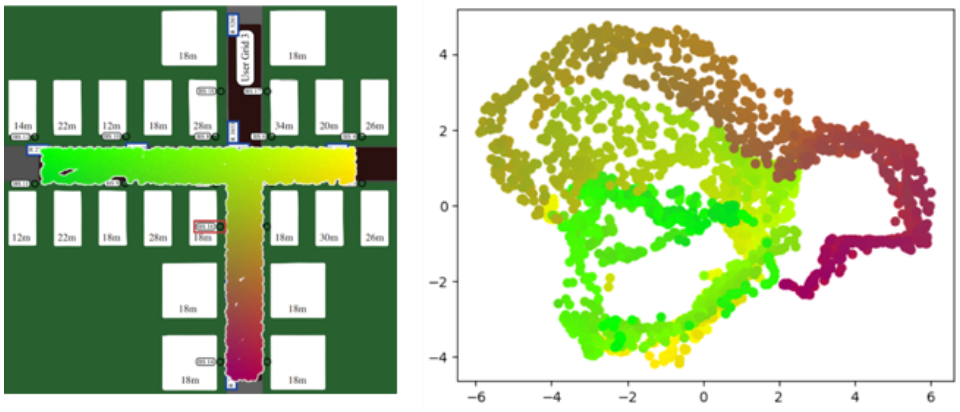

Figure 3-6: Example of urban environment (left) and corresponding channel chart (right). Each coloured dot corresponds to a location where the channel has been measured.

3.4.2 Interference management in cell free massive MIMO

In denser deployments, cell-free massive MIMO systems outperform cellular small cell systems in terms of spectral efficiency [BS20a]. Along with this user-centric communication architecture, challenges arise in interference management and access point (AP) allocation for the users under constrained fronthaul link capacity. Most of the current algorithms, [BS20b, LZJ+20, SPW18, BDF+21] to allocate pilots and APs to the users operate on the assumption that accurate large-scale fading coefficients to all the APs in the system are available either at the user or at the CPU. Acquiring this statistic accurately for a large number of APs in a dense deployment is challenging and may not even be realistic under 5G new radio (NR) signalling schemes. Furthermore, in most

of the literature, cell free systems are considered as a stand-alone system. Therefore, optimal selection of users to be served non-orthogonally on the same resource block in a user centric manner is yet to be explored.

For pilot allocation, in [LZJ+20] authors have proposed a graph colouring-based algorithm, where contaminating users which share the same pilot sequence are represented as connected nodes in a graph. [SPW18] proposes Hungarian method-based pilot allocation methods, where one is maximizing system throughput while other is maximizing the fairness across users. In [BS20b] authors present a scalable AP selection and pilot allocation method based on large scale fading. Here scalability arises from the fact that only neighbouring APs to the user are selected.

As a communication network can be represented by means of a heterogenous graph, we aim to leverage the novel advancements in graph representation learning [WPC+21] for interference management and AP selection under sparse availability of inaccurate large-scale fading coefficients. Furthermore, the idea of optimal user selection for non-orthogonal communication needs to be considered as cell free systems co-exist with a multiplexing scheme like OFDMA.

These novel ML/AI-based interference management techniques would mainly support the interacting and cooperative mobile robots use case where new cell-free massive MIMO architectures will be used to manage a cluster of drones over a 6G network aided with novel ML/AI-based interference management, pilot allocation and AP selection algorithms.

3.4.3 Radio resource allocation cell-free massive MIMO

Radio resource allocation in cellular and cell-free MIMO networks improves system performance by efficient utilisation of available resources such as power and bandwidth. Conventionally, the radio resource allocation problems are solved using optimization or heuristics-based methods, considering channel state information (CSI) and QoS requirements of the users. These methods have several challenges, such as high computational complexity and requiring precise CSI, resulting in sub-optimal solutions in complex and non-convex problems, lack of flexibility and parameter sensitivity, and inaccuracy of the model-based resource allocation methods (due to channel modelling issues and hardware impairments) [HHH+20]. ML approaches can be used to overcome these algorithm deficit or model deficit challenges of the conventional resource allocation techniques. Some of the identified problem scenarios are:

1) Joint power control and fronthaul signalling capacity allocation in cell-free MIMO uplink scenario. In cell-free MIMO, the network performance is determined by different levels of coordination among access points and CPU which result in different processing requirements in those components and different fronthaul signalling requirements. Proper utilisation of the transmit power of the users and fronthaul capacity between access points and the CPU enable performance improvement of the network.

2) Joint power control, beamforming, and antenna subarray selection in communication systems approaching the upper mmWave range to improve the spectral efficiency/energy efficiency.

3) Resource allocation in massive MIMO systems with hardware impairments.

In the literature, deep learning techniques have been proposed as a low complexity solution to the complex optimisation algorithms. Furthermore, DRL techniques enable learning of control actions with trial and error based on observed system parameters and dynamic system behaviour over time. They can be used in instances where prior knowledge of the network, parameters and users is not available and needs to be predicted and resource allocation decisions to be made [HHH+20].

Several studies have proposed deep learning-based power control for cellular and cell-free massive MIMO systems [ZND20, SZD18, DZB+19, LSY+20, VNB+20]. Most of the existing studies focus on a supervised learning approach, where a deep neural network (DNN) is trained to learn the mapping between the inputs (user locations or channel statistics) and the optimal power allocations obtained by an optimisation algorithm. The unsupervised learning algorithm proposed in [LSY+20] for multi-user interference channel power control problem eliminates the need of knowing the optimal power allocations during model training, hence has a simpler and flexible model training stage. Unsupervised learning-based uplink power control in a cell-free massive MIMO network for minimum user rate maximisation is presented in [RMR+20], where the DNN is directly trained to achieve the desired objective using the large-scale channel coefficients as the inputs to the model.

A distributed dynamic power allocation scheme based on model-free DRL is proposed in [NG19]. Each transmitter adapts its transmit power based on CSI and QoS information obtained from its neighbours. It is shown that the proposed deep Q-learning approach gives near optimal power allocation in real time to achieve the weighted sum-rate maximisation objective. Authors in [MEA20] have proposed a DRL framework to solve the non-convex problem of SINR maximisation by joint beamforming, power control, and interference coordination. Simulations performed at sub-6 GHz and mmWave frequencies have shown, see Figure 3-7 to exhibit acceptable performance compared to existing standards and run time reductions.

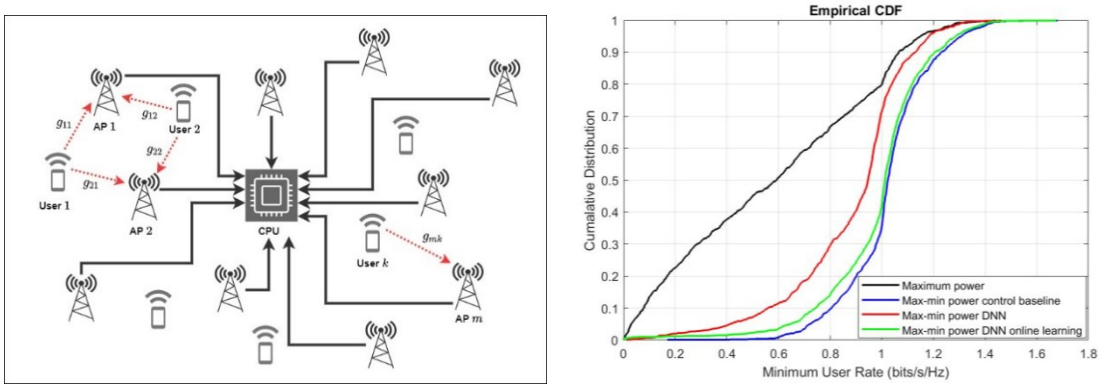

Figure 3-7: The cell-free MIMO system model (left) and minimum user rate performance of deep learning-based power control proposed in [RMR+20] in comparison with maximum power transmission and optimization-based baseline power control algorithm (right)

For the joint power control and fronthaul signalling capacity allocation problem in cell-free MIMO, deep learning-based joint power control and fronthaul capacity allocation is proposed, where a DNN is trained to learn the optimal user power allocations and optimal fronthaul capacity allocation for CSI and data transmission from the access points to the CPU in an unsupervised manner, in order to achieve maximum sum-rate or user fairness.

RRM becomes more challenging in upper mmWave networks due to the increased system complexity and high dimensionality of the resource allocation problems due to large number of antenna elements in the transceivers. DRL or deep learning-based approaches are to be explored for solving those complexity issues joint optimisation problems in upper mmWave networks.

Massive MIMO power control in non-ideal system setups, such as in scenarios with transceiver hardware impairments or ,where perfect CSI not available are also planned to be investigated, to understand the potential of deep learning-based resource allocation approaches in practical applications.

The above-mentioned ML/AI-based resource allocation approaches would be enablers for the interacting and cooperative mobile robots use case where new cell-free massive MIMO architectures will be used to manage a cluster of drones over a 6G network along with novel ML/AI-based resource management, link adaptation and AP selection algorithms. In addition to achieving the traditional communication KPIs, ML-related KPIs such as complexity gain, flexibility, convergence and mobility support will be considered when determining the performance of the proposed algorithms.

3.4.4 Data importance-aware RRM

The problem consists in prioritising data point transmissions per a *data importance* (or *data quality*) criterion, which is expected to impact the way available radio resources, such as time, frequency and power are allocated to wireless devices contributing data. Such prioritisation is challenging, due to the joint presence of channel uncertainty (interference, noise, channel model mismatch) and data uncertainty (measured by entropy, or, simply the distance from the current model decision boundary) and needs to be addressed when both centralised and FL setups are considered.

Active learning consists in selecting important samples from a large unlabelled dataset for labelling (by querying an oracle) to accelerate model training with a labelling budget (see [WLZ+19] and references therein); direct advantages that one may immediately think of are data economy/ frugality, as only smaller, albeit important data samples are dispatched and processed for learning purposes. Additionally, such dataset compression also translates to less storage space occupied and higher communication efficiency. A widely adopted measure of (data) importance is uncertainty; specifically, a data sample is more uncertain if it is less confidently predicted by the current model. A commonly used uncertainty measure is entropy, a notion from information theory. As its evaluation is complex, a heuristic but simple alternative is the distance of a data sample from the decision boundaries of the current model; the farther this data point lies from the current model decision boundary, the lower its importance is.

A basic challenge of applying this concept to a wireless communications system is that, as compared to active learning, which is characterised by data uncertainty, RRM implemented for the needs of ML model learning is characterised by joint data and channel uncertainty due to wireless channel volatility.

To tackle the joint data and channel uncertainty challenge, [LZZ+19] proposed a solution based on an importance-aware retransmission scheme for data acquisition. The scenario consists in an edge learning system, where a classifier is trained at the edge server based on SVMs, with data collected from distributed edge devices. The problem consists in that acquisition of high-dimensional training data samples is bandwidth-consuming and relies on a noisy data channel. On the other hand, a low-rate reliable channel can be, instead, allocated for accurately transmitting small-size labels. Such resource allocation (e.g., time) across data contributing devices needs to be carefully designed, as mismatches between labels and noisy data samples may lead to an incorrectly learned model. To solve this issue, [LZZ+19] suggested an importance-aware retransmission protocol with coherent combining to enhance data quality. Other works, such as [ZLT+21] suggest measuring the significance of data samples by data diversity, which is defined as the difference between data samples; [ZLT+21] then proposes a joint data-and-channel diversity aware multiuser scheduling algorithm to optimise communication resource efficiency without limiting the model's inferencing capability.

Regarding the applicability of the data importance-aware RRM concept to FL setups, [ZLD+20] proposed a scheduling policy to exploit both diversity in multiuser channels and diversity in the "importance" of the edge devices' learning updates. The basic principle is that each device

communicating its locally updated model parameters to an edge server, also provides its own importance indicator report. The collected reports, together with uplink CSI are then used as input to a centralised scheduler which selects local model parameters needed to update (and then share) the global model. The topic of scheduling for cellular edge FL with importance and channel awareness has been covered by [RHW+20], as well.

The above-mentioned state-of-the-art works have initially covered the topic, however, mostly focusing on time as a resource to be optimised per a data significance awareness criterion. Also, various criteria may be thought of to evaluate data significance, based on which RRM schemes can be designed (refer to Figure 3-8). Some possible metrics are the following:

1. Data Age-of-Information (AoI), a time-evolving measure which characterises information freshness at the receiver (e.g., at the cloud edge where an ML model is updated). The AoI at a given time instant is defined as the time difference between the focused timestamp and the time at which the observed state (or data packet) was generated. As a general principle, devices providing more "fresh" learning data may be scheduled over other devices in order to keep the trained model relevant.
2. Correlations among datasets coming from various devices experiencing similar context (e.g., video feeds coming from cameras with similar Field of View (FoV)). Depending on the scenario, groups of devices providing highly correlated learning data sets may be scheduled over others depending on the inferencing task to be performed.
3. Possibly other metrics (e.g., data contributing device inactivity time etc.).

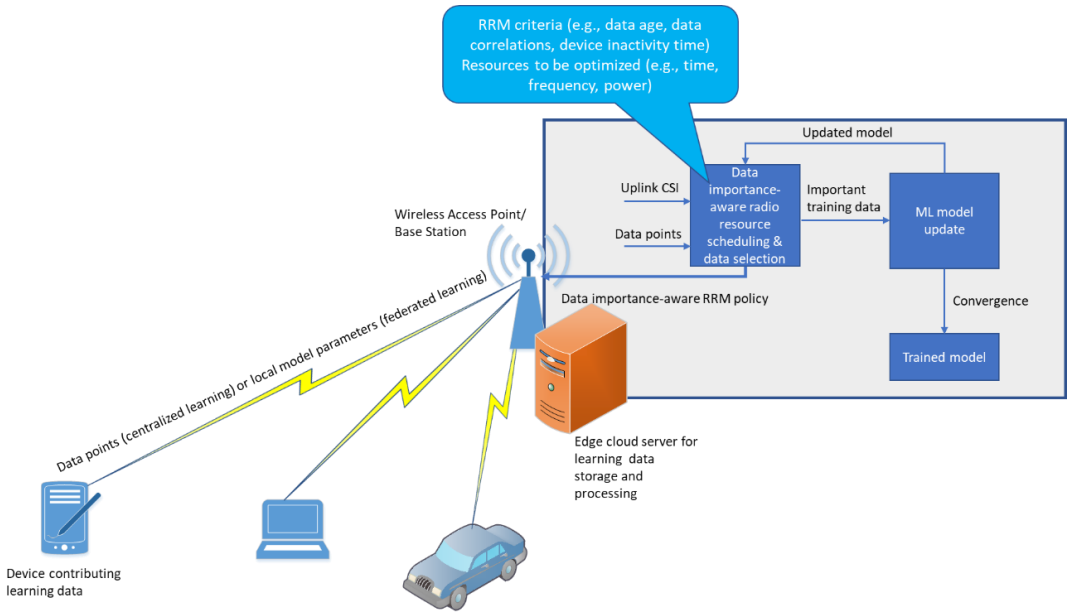

Figure 3-8: Data importance-aware RRM and data contributing device scheduling

It should be noted that, depending on the data significance criterion considered for each scenario, the structured learning database may be maintained/updated likewise (e.g., dispose data points the AoI of which is over a threshold, i.e., "stale" data). On top of the criteria to be used for data importance evaluation, technical solutions may also focus on the radio resource domains (i.e., beyond time) that can be exploited for data importance-aware RRM. Examples refer to frequency, power and antennas.

Data importance-aware RRM is expected to be a feature of a dynamically changing environment in which AI/ML-based decisions will need to be taken considering limited compute and data

storage capabilities, as in private local networks. For that reason, solutions are expected to address latency, signalling overhead and energy efficiency requirements of 6G use cases, such as the "Flexible manufacturing" one. Additional ML-related KPIs to be considered (beyond the analysis in [D1.2]) are the ones of model convergence speed and data quality, whereas, new relevant KVIs are the ones of generalisability and distributed learning with frugal AI.

3.4.5 AI for distributed massive MIMO architectures

Massive MIMO is a key technology component for high-performing wireless networks, which provides high beamforming gain and leads to increased spectral and energy efficiency. Recently, distributed massive MIMO (D-MIMO) systems are investigated heavily as a potential MIMO architecture to meet the requirements of envisioned use cases, KPIs and KVIs in the 6G timeframe [D1.2]. For example, use cases of immersive communication, networks of networks or merged cyber-physical worlds and flexible manufacturing may all imply a denser network infrastructure. The massive Distributed MIMO architectures, also called cell-less systems are one possible solution for network densification. These systems implement user-centric transmission to cope with inter-cell interference and provide additional spatial diversity. The increasing number of antennas in the system heavily increases the complexity of the optimisation problem to solve (e.g., how to select pre-coding and power allocation schemes, selecting APs/antennas, pilot assignments, etc.). D-MIMO poses further implementation issues to become an efficient scalable alternative to existing solutions due to the distributed nature of system components: antenna elements, compute resources, as well as channel measurements. Therefore, this complex optimisation task should be solved considering a lot of practical constraints, such as limited fronthaul capacity or limited communication exchange between APs (e.g., exchange of CSIs) or at a limited set of radio transceiver chains, or limited power consumption. Although, this task can be presented as a large joint optimisation problem, it is reasonable from practical applicability point of view to decompose it into sub-problems and use AI components to solve these sub-tasks. The inter-play of such multiple AI components raises some interesting further research questions.

Distributed MIMO has a number of architectural variants depending on which processing functions are centralized and which ones are distributed, ranging from a fully centralized to a fully distributed architecture and anything in-between. For example, in a fully centralized approach, all signal processing, pre-coding, scheduling, UE-AP assignment are done centrally, which may result in a close to global optimal solution but may raise scalability, fronthaul bottleneck and other issues. In the centralized case, the AI components would also need to be more of centralized type. In a more distributed approach, where processing functions are shared between a central entity and local APs, the AI-based solutions would need to follow a distributed approach as well. See Figure 3-9 for an illustration of a possible D-MIMO architecture combined with distributed AI functions in APs and CPU. For example, AI logic in the local APs and central unit may cooperatively solve channel prediction, UE-AP assignment, pre-coding or other resource assignment issues. The multiple AI components need to act cooperatively to avoid conflicting decisions and achieve an overall optimal or sub-optimal solution. Therefore, it is important to study the different AI solutions in the context of different D-MIMO architecture realizations.

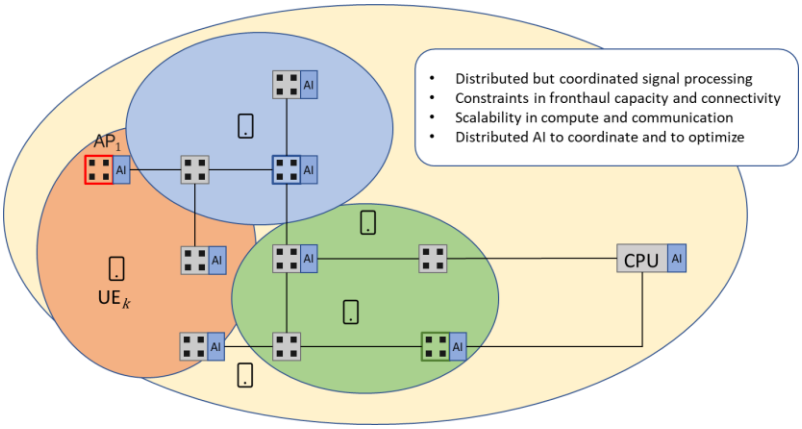

Figure 3-9: Distributed MIMO architecture combined with distributed AI support

There are opportunities to handle the extra complexity of a transmission problem arising from the massive number of antennas and from the distributed nature of the architecture with AI intelligence, where the role of AI could be in (1) solving these large optimisation problems in parts or in complete or (2) mitigating the effect of some practical hardware and implementation constraints (e.g., lack of full CSI knowledge or finite fronthaul or hardware resource), (3) providing spatial and temporal predictions for dynamic measures based on distributed AI. For example, knowing the large scale pathloss to the set of APs is a required information in many D-MIMO processing algorithms. Using prediction either in time or across antennas can help to maintain this information without continuous measurements.

Another important AI aspect in the D-MIMO context but also in the more general AI for physical layer context is the continuous and online learning. In online learning [BKP19] the data is arriving continuously, and the model is updated online and being applied in parallel to learning. Continuous learning is when new information is learnt without forgetting previous knowledge. Typical neural networks suffer from catastrophic forget [KPR17] when previous knowledge is forgotten as new data and new knowledge is obtained. Meta-learning techniques are possible solutions to avoid catastrophic forgetting [FAL17]. As in a radio environment, the channel patterns, e.g., the antenna relations, antenna cross-correlations in a D-MIMO system may change constantly, it is important to employ AI algorithms that can quickly learn the current patterns without forgetting the more generalizable commonalities in the different radio environment datasets. The architectural implications supporting online learning should be an important system design aspect.

Cell-free architectures are expected to be highly relevant in providing reliable, high density and high capacity deployments, where interference management and coordinated transmissions are essential for the correct operation of the system. Since multiple APs take part in serving a UE in cell-free systems and the connectivity is continuous to multiple APs, these systems can provide the extra reliability and the seamless mobility that are required for mission critical applications to be encountered in use cases such as the "Flexible manufacturing" or the "Interacting and cooperative mobile robots" use cases. Using AI/ML techniques to manage the radio resources and coordinate transmissions, resource allocations in such a complex, distributed system seem to be a natural and necessary choice to fulfil the communication KPIs and to serve these use cases.

3.4.6 Model predictive control for MIMO antenna systems

In systems theory, Model Predictive Control (MPC) [CA13] covers an ample range of control methods that make explicit use of a model of the process to obtain the control signal by

minimizing an objective function. For MIMO antenna system control, the general goal of MPC is to assist complex rule-based systems by learning a close-to-optimal control for example of power and bandwidth allocation, precoding, etc.,so that the final control can be refined by a traditional rule-based solution, linear approximation, or even a fast reinforcement learning method on the fly [FDA17].

In the SoTA, MPC approaches towards radio telecommunication appeared early. In [XMT17], the congestion control task is considered, models are trained to combine multiple linear dynamic capacity allocation schemes (operating points) and NS2 simulations [TH09] are used for evaluation. In [BB12], the task of determining when sensors should communicate to maximize their battery life is considered. For MPC, two variants are considered: they either predict a series of next control steps (explicit MPC) or the first step only (implicit MPC).

The main question regarding the use of MPC for 6G MIMO antenna control is whether there are data of sufficiently long time interval that includes sufficiently wide variety of radio control settings to train and evaluate MPC. 4G wireless service (Long Term Evolution – LTE) data is available in the largest amounts, which may already be complex enough with 100s of setting parameters to evaluate control approaches.

The most natural machine learning approach for MIMO control is reinforcement learning (RL). Q-learning [WD92] is a model-free general approach that requires no knowledge on the system to be controlled, rather just the reward (Q) function. However, as pointed out in [GLSL16], Q-learning needs too many samples for continuous action problems such as antenna control. In such cases, model-free learning must be combined with model-based learning that involves expert knowledge in the learning process. Model-based RL relies on known or linearly approximated system behaviour. By RL, we have to enforce a trust region not to deviate too much from the region in which samples were generated. Here, RL serves as a fast model for compute performance, while safety can be decoupled under reasonable conditions in an optimization framework by maintaining another, traditional model of the system [AGST13].

The technology gap in modelling approaches, especially with RL, is the lack of quality guarantees, which would be crucial for the service operation. We have to distinguish the provenance of modelling approaches whether they are online trained on the fly, or built and evaluated as a batch on previously collected training data. In the latter case, we have the opportunity to experiment with all reasonable model settings and evaluate the model behaviour under all foreseeable events. However, if we learn models on the fly, which RL normally does, the models are immediately deployed without a human in the loop. In [TL20], batch and online ML are considered as complementary approaches. Online ML has the advantage to quickly adapt to changes in the environment. Optimality strategies (competitiveness) and guarantees of batch models compared to online ML are described in the paper.

Overall, the foreseen solution direction is to evaluate ML approaches on long time range and wide area data of radio control and network performance. Applications include the control of power allocation, or precoding (possibly in the presence of hardware impairments), or fronthaul signaling bandwidth allocation in distributed MIMO systems with limited observability of channel state information.

We aim to test and compare batch and online ML and RL methods. Rule and physical model-based approaches are required at two places. First, Q-learning can be combined with models; second, approximate and heuristic MPC results can be corrected by such approaches, as in [GLSL16]. The use of RL, at least in its original concept, does not clearly separate the training and production stages: it is assumed that the system is integrated into the production environment and learns with real data taken from that environment. Since suboptimal responses can still occur

in the production stage, a possible approach to handle such situations could involve performance guarantees and deploy rule-based safety control components in a real production environment.

4 Motivation and gaps for in-network learning methods & algorithms

The purpose of this chapter is to identify gaps and key issues of implementing *in-network learning methodologies and algorithms*. As part of gap analysis, several academic literature works have been reviewed together with relevant work in wireless communications standards, such as 3GPP [22.874] and the ITU (Focus Group on ML for Future Networks including 5G - [ITU20]). An overview of the covered technical areas and their interconnections appears in Figure 4-1, also highlighting relation to the overall Hexa-X research challenges of Sustainability, Trustworthiness and Connecting Intelligence.

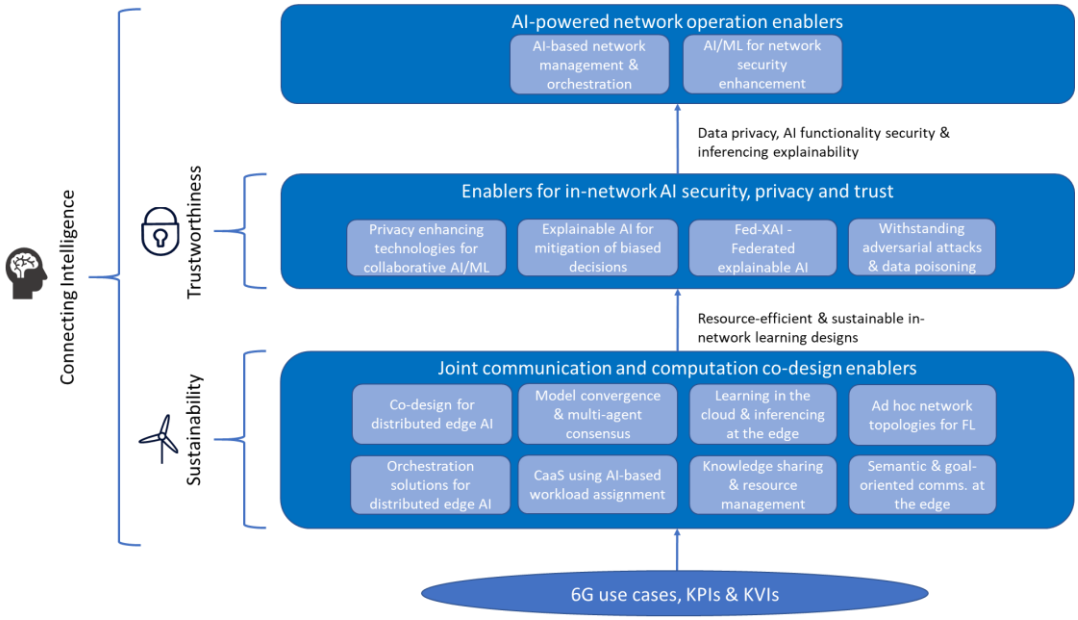

Figure 4-1: General overview of technical areas on in-network learning; each area involves multiple technological enablers

With regards to the relevance this chapter's technical enablers have with the identified use cases, as listed in Section 2.6 of this deliverable and per their detailed description in [D1.2], Figure 4-2 explains such relevance-based pairing. Based on it, Figure 4-3 aims to further define KPIs and KVIs sourced from [D1.2] along with introducing new ones; KPIs and KVIs of specific importance to this chapter's topics appear in the highlighted rectangular schemes of the figure.

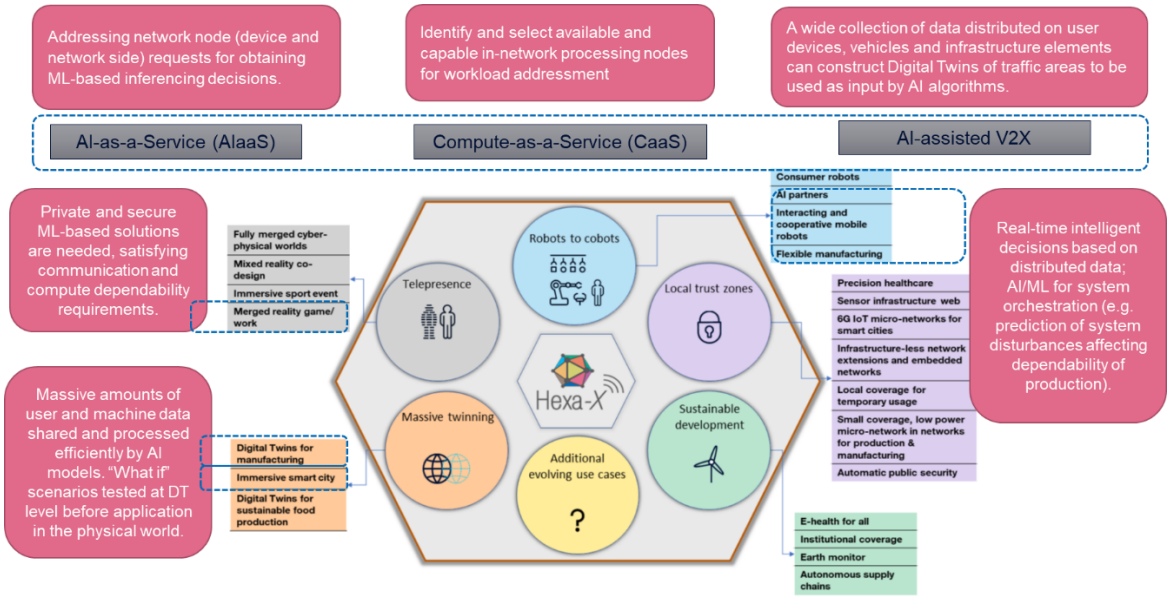

Figure 4-2: Tailoring technology enablers on in-network learning to identified Hexa-X use cases, based on [D1.2]

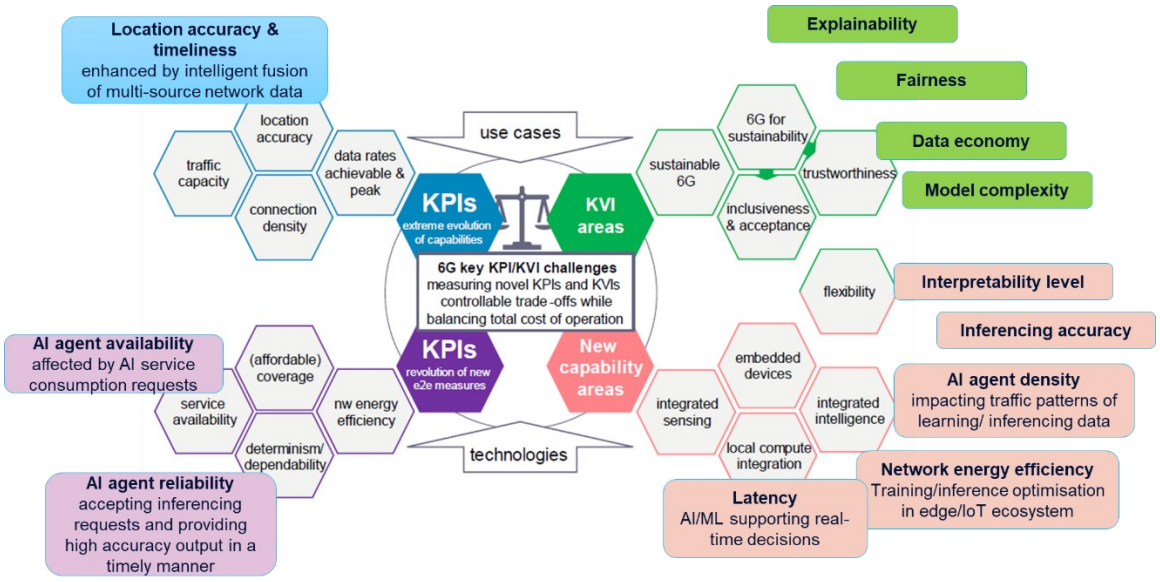

Figure 4-3: Proposed expansion of Hexa-X sets of KPIs and KVIs with indicators relevant to in-network learning

The chapter starts off by covering the topic of *joint communication and computation co-design as an enabler for distributed learning* in Section 4.1. A first approach of such co-design is detailed targeting efficiency improvements of distributed edge AI. The incurred challenges in model convergence and multi-agent consensus for distributed ML setups are also elaborated, along with the concept of learning in the cloud and inferencing at the edge. The advantages and roadblocks of realising ad hoc network topologies for FL are analysed, along with orchestration solutions for distributed edge AI. AI-based workload assignment is proposed for trustworthy and sustainable CaaS operation, whereas knowledge-sharing mechanisms are deemed as useful to support the AI functionality of the network. Semantic and goal-oriented communications are suggested to enhance AI/ML communication and energy efficiency at the network's edge.

Section 4.2 details *enablers for in-network AI privacy, security and trust*. Privacy-enhancing technologies for collaborative AI/ML are reviewed, along with the role of explainable AI for mitigation of biased decisions and the design of FL systems able to explain themselves. The problem of data poisoning, i.e., an attack type on AI functionality consisting in perturbation of training data (or model parameters) aiming to lead to false generalisations, is also elaborated and adversarial attack mitigation means are suggested.

Section 4.3 focuses on *AI-powered network operation*. In further detail, in this section, the advantages and challenges of AI-based management and orchestration for behaviour-driven adaptation are explained, along with the role AI/ ML can play to shield the network from security attacks, based on data-driven anomaly detection-based solutions.

4.1 Joint communication and computation co-design as enabler for distributed learning

One of the key innovations in future wireless networks will be the combined design of communication and computation due to the growing necessity of having more and more computing resources at multiple nodes of the network, in particular at its edge. Part of those computational resources will be dedicated to the intelligence of the network, i.e., ML-enabled computing nodes will allow for a distributed intelligence in the network. With scope on distributed learning and AI, this section covers some of the gaps related to joint communication and computation designs envisioned for 6G systems.

4.1.1 Communication and computation co-design for improved efficiency of distributed edge AI

AI applications and computations are traditionally executed at a centralised resource pool, due to the need for acquisition of large learning datasets, which, in its turn, implies heavy computational processing needs. The increasing adoption of intelligent components in mobile gadgets, IoT devices and other machines indicates a shift of having more and more intelligence deployed at the edge of wireless networks. Further performance and functional requirements, like constraints on processing latency or data locality also point towards distributed deployments. This trend motivates the need to develop intelligent edge functionalities for an AI-native 6G network, capable of supporting intelligent systems for both network and external/ third-party applications. Supporting capabilities include novel AI-native communication solutions, as well as compute functions in edge equipment. A new research area, called *edge AI*, is emerging to combine the disciplines of wireless communication and AI, where the main challenges are to overcome the constraint of IoT devices of limited computing capabilities, take the special communication patterns of distributed intelligence into account and ensure data trustworthiness. In such systems, learning from distributed data and communicating between the edge server and devices are two critical and coupled aspects.

Some AI architectures are well positioned to exploit the massive data distributed over wireless devices, but their design must be harmonised jointly with 6G network features to unlock their full potential. Edge distributed NN frameworks [ZCL+19a] are constructed to scale out a single model by splitting along input data or model segments with supervised learning. Federated Learning (FL) [KM+21] is a widely applied and researched architecture, which poses significant workload on wireless edge, therefore, efficient network co-design solutions are needed for resource optimisation, as well as increased privacy. The field of Sparse NNs is an actively researched area of NN architectures [LSZ+19]; the sparse activity property means that only a small fraction of

neurons is active at any time, whereas the sparse connectivity property means that each neuron is only connected to a limited number of other neurons. Such sparsity properties may be useful from the point of communications and computation efficiency, as lower workloads are transferred over the air. Devices that are interacting with the physical world to do joint tasks scale better in multi-agent architectures, like collaborative Multi-Agent Reinforcement Learning (MARL) [HKT19]. In addition to alignment of 6G networks with the above existing AI frameworks, research in this field should also cover emerging AI architectures, such as biologically inspired NNs and, in particular, Spiking NNs (SNNs) [TGK+19], since they pose radically different requirements on the communication infrastructure. SNNs imitate the behaviour of biological neurons and are often mentioned as the next AI generation. Similarly to a biological network, a distributed set of devices would communicate with spikes (discrete events that take place at points in time, rather than continuous values) over wireless channels. Due to the nature of spike-based communication and its communication features, special requirements need to be fulfilled to effectively support distributed neuromorphic-based AI applications. In a SNN only active neurons transmit information; such a feature is very energy-efficient, but it must be supported by similarly efficient communications, like lightweight access control. The exact time of a spike conveys important information, so the communication medium must keep this relation with high fidelity. Another aspect is that in neuromorphic systems spike transmissions are highly localised, which may benefit from network features for local communications.

Edge distributed NNs can be categorised as model split and data parallel architectures. Data parallel NNs are used in, e.g., FL, when local training data are naturally available by wireless devices, and the aggregated/ federated model can be trained collaboratively by the devices sharing their local models' gradients or weights, without exposing any local data. However, the model updates to be shared can be in the form of large payloads, which motivates recent and future research to increase efficiency with, e.g., over-the-air compute for reusing time-frequency resources or decentralised training. Another example is hierarchical fog learning which benefits from dynamically formed local communication groups managed by the network. Wireless resource-aware scheduling of transmissions, model compression by pruning or gradient compression by quantisation can also efficiently reduce data volume [DZF+20]. Some of these techniques can be applied for model-split networks and feature distributed NNs.

While edge distributed NN models are mostly of fixed structure, many AI applications require a more flexible and scalable architecture for interacting with the environment. MARL provides a solution for dynamic scalability with localised decisions of interacting agents. The challenge of scalability and efficiency motivated research in MARL architectures [KMH+19], [PSB+19] on what/how/when to share from locally available data among agents.

Efficient training and execution of these applications in a wireless environment requires well-coordinated operation with the underlying communication network. 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. While applying these concepts, optimisation of the AI models can tolerate a lack of precision and errors, which allows the design of more resource-efficient transmission modes. A toolset for communication-efficient inference 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.

Use cases with highly distributed data and delay constraints benefit from the joint design of communication and computing. The "Interacting and cooperative mobile robots" use case requires real-time intelligent decisions based on distributed data, with resource efficient data and model sharing. In the use case of "Immersive smart cities", a huge amount of data will be distributed on thousands or millions of devices, all of which is not feasible to be shared due to communication

capacity, privacy, complexity and other reasons. By making available this data from many devices, machines and infrastructure elements in a trustworthy, energy and resource efficient way for distributed AI applications, it will be possible to make decisions with higher accuracy and lower inferencing latency as main KPIs. High device (and AI agent) density is also an important KPI to measure the efficiency of the proposed solutions. Such a system contributes to the key value area of sustainability by targeting data economy.

4.1.2 Model convergence and multi-agent consensus in distributed ML

To design an efficient decentralised ML system in radio networks, it is imperative to understand the interplay between data compression, resource allocation, and overall performance [SZHT07]. There is a gap between performance, communication, and complexity. For example, combining information from neighbouring nodes via in-network signal processing can improve reliability and reduce the amount of traffic on the network. On the other hand, the exchange of additional information could potentially yield better decisions [SW95].

Another gap in decentralised or fully distributed ML solutions is the lack of guarantee of convergence compared to a centralised solution, both in theory and in practice. For example, fundamental performance limitations are known for intrinsic average consensus problems in open multi-agent systems, which are systems subject to frequent arrivals and departures of agents [MH19]. The bounds depend on the properties of the model defining the interactions between the agents and instantiate that result for all-to-one and one-to-one interaction models and no general solution exists.

Successful approaches for distributed ML rely on gradient descent algorithms. Positive results and experiments are also present in the literature, for example for low rank approximation tasks [HBK+16], which open opportunities for building recommender systems. New variants of sampling strategies in coordinate descent are analysed in [GOP19], who analyse the convergence rate of the random reshuffling method for minimising a finite sum of convex component functions and give high probability convergence rate estimates. Coordinate descent approaches might, yet, possess less explored potential for use in decentralised ML [HBK+16]. Communication patterns for distributed learning are analysed by [RTM+19], who identify stragglers and communication overhead and mitigate by setting deadlines and quantising local model share. As a new approach, the aim is to investigate optimal messaging for distributed training of NNs by identifying and then sharing important activations in deep neural models by techniques used in adversarial attacks [ZSM+18], which might not only reduce communication, but more importantly also speed up convergence [S15].

4.1.3 Distributed learning – learning in the cloud and inferencing at the edge

Reduced capabilities devices, such as seen in IoT use-cases, pose challenges to the training of models at the edge. Battery lifetime constraints, as well as limited computation capabilities, hinder the training of existing full-size models in a sense that lacks adaptation to changes in the environment and considering the individual characteristics of the involved devices.

While running models – or inferencing – on constrained devices seems realistic, thanks to the availability of low-power Neural Processing Units (NPUs) [RMJ+20] and current research on low-complexity models [LDD+20] [LDL+21] [NML+18], training those models on the same hardware is likely unfeasible or too energy-consuming [RMJ+20]. Distributed learning schemes, where computation-intensive training is processed in the cloud and only small adjustments are handled at the edge devices are, therefore, of interest.

Supporting such delegation schemes requires to fill the following gaps: identification of relevant data for the training process of the models, procedures for requesting and pushing training data and model parameters updates to/from the devices and signalling channels to support them.

Such strategies pertain to the concept of Edge Intelligence described in [ZCL+19a], more specifically level 3 and 4 of their classification, i.e., "on-device inference and cloud training" and "cloud–edge co-training and inference" respectively. Knowledge transfer learning techniques, and, more specifically, teacher-student schemes, are promising approaches to the problems of adapting a model to a specific task [SBL+18], such as the equalisation in presence of hardware impairments, and the parameterisation of low-complexity edge models using complex models trained in the cloud. Moreover, low-complexity, explainable models, such as [LDD+20], [LDL+21], are interesting edge (student) models to investigate those learning strategies, as they meet both the complexity requirements of IoT devices and can be analytically parameterised with provable equivalence to traditional algorithms.

Envisioned promising approaches related to the transfer learning and knowledge distillation technique include studying the correlation of student and teacher models' parameters to derive on-device update strategies that do not require retraining models in the cloud. Possible techniques to evaluate those correlations involve non-linear Principal Component Analysis (PCA) approaches using NNs, e.g. [Kra91]. Weight-freezing strategies, such as [ISM01], are also interesting candidates for incremental re-training of edge models, as they reduce the number of trained parameters and, therefore, the computation toll on the edge device.

Such delegation schemes between edge and cloud are deemed to improve conventional communication KPIs such as energy efficiency (notably regarding battery lifetime of constrained edge devices) and adaptability to environment changes. Low complexity and explainability of edge AI models are expected to reduce training complexity and are, therefore, KVIs of great interest.

4.1.4 Ad hoc network topologies for federated learning

Federated learning (FL) in a multi-agent wireless network environment prone to failures requires resource replication and overlay network topologies that are robust to failures and inconsistent views of the entire network, for example, when certain AI agents are unaware of the availability or unavailability of certain others. The most widely used overlay solution for balancing the computational load with approximate knowledge on the availability of other peers is consistent hashing [LJC+00], which has a variety of applications in distributed key-value stores, peer-to-peer networking, and ad hoc routing.

A gap consists in the absence of failure-resilient overlay network structures that take away uncertainties in an abstract network layer. Perhaps the first geo-aware overlay network is proposed in [LJC+00]. In [BGL03], optimal overlay structures were considered, based on hypercubic networks, which minimise the number of peers in direct contact and the hop length between the resources. Future work may consider local close cooperation group formation for geo-aware FL by relying on the geographic routing ideas.

Currently, the research community is actively seeking how to combine results on network topologies and resource allocation by exploiting special network capabilities including downlink multicast and uplink compute-over-the-air with analog operations [ZWH20]. There is a need to further investigate multi-layer hybrid learning frameworks [HBA+20]. Such networks consist of heterogeneous devices with various proximities to account for the topology structures of the local networks. Among the heterogeneous nodes, cooperative learning functionalities can be provided through device-to-device communications, using location-aware distributed topologies at scale--

-since such topologies provide logarithmic performance as the function of the size, the approaches might be scalable to millions of devices [LJC+00].

Another goal of upcoming work would be to combine traditional overlay network technologies with computation-aware communication techniques that are recently starting to draw attention. Computation-in-the-air [ZWH20] assumes analog operations in that Distributed Stochastic Gradient Descent (DSGD) model updates are simultaneously transmitted by devices over broadband channels and aggregated over-the-air by exploiting the superposition property of a multi-access channel. In such setups, interference is harnessed to provide fast implementation of the model aggregation.

4.1.5 Orchestration solutions for distributed edge AI

This section presents an overview of current solutions for the orchestration of algorithms in support of distributed edge AI, and the analysis of potential gaps to be filled in a 6G network in this area. Here, the focus is mostly on the orchestration of AI tasks and sharing of AI models. The applicability of such solutions to the orchestration of 6G networks and services as a whole is further discussed in the context of WP6.

State of the art solutions [FPE17] for orchestration and slicing in the 5G networks are mainly based on AI/ML techniques adopting a centralised approach. Such techniques are applied to several optimisation areas. For example, the prediction of high-level and generalised requirements such as latency, maximum number of users, bandwidth, mobility, and others, can be applied to enhance the resource orchestration for heterogeneous types of services, spanning from mobile broadband communication to low-latency and critical communications. AI and ML techniques applied to network slices optimisation, like the one proposed in [GGD+19], can improve efficiency of the network resource utilisation exploiting the concept of resource elasticity at the orchestration level. Other solutions are addressing more specific targets. For example, in [CHW+20] ML techniques are applied to network slicing strategies for vehicular networks, driving the allocation of network resources on the basis of traffic predictions.

Despite the promising results, the centralised approach adopted in the aforementioned solutions presents some limitations, mostly related to data privacy, load of traffic to move the data towards the cloud, energy efficiency and latency. The last point is especially critical for orchestration actions related to service re-configurations that would need to be applied in very short periods and, thus, would benefit from reaction decisions taken and enforced in near real-time. In this direction, the distribution of AI/ML processing towards the edge of the network is gaining particular attention especially in IoT, industrial and autonomous driving contexts. Considering the requirements of new 6G use cases in terms of extreme service dynamicity, which are lowering the acceptable timing of orchestration reactions, this trend will continuously accentuate towards solutions bringing not only the inference, but also the training at the edge, up to the device level [PBC+20].

At the current state, orchestration solutions predict only classical high-level requirements, such as latency, maximum number of users, bandwidth, geographical constraints, and alike. With the birth of new and more reliable technologies, but also accounting for concerns related to data sensitivity and communication system sustainability, new requirements, such as extreme low latency, one Tbps data rates, privacy, energy efficiency, and AI trustworthiness must be taken into account. This new set of requirements [YWH+21] opens up new frontiers in the design of 6G resources management, controlling, and orchestration bringing to a reconsideration of the existing architectural solutions and standards in the orchestration domain.

For this reason, the orchestrator has to find the most suitable solution in terms of what, where, how, and when an Edge DNN should be deployed and distributed in the edge network, in

particular among the edge and the devices. Different Edge DNN architectural solutions have been proposed to distribute the AI compute workloads among the devices, the edge, and in some cases in the cloud. In particular, in [ZCL+19a], the solutions span from the device-based model, where the inference is made by the device itself, to the edge-cloud mode where the DNN model is executed by edge and cloud. It is quite obvious that a single architectural solution is not feasible for all use cases and scenarios, but rather strictly depends on the high-level application requirements and constrained by the edge resources availability. In the context of 6G, high-level requirements of applications running at the edge of the network, such as latency, security, privacy, reliability, energy efficiency, accuracy of DNN solution, costs, and others are very dissimilar and, in some cases, even conflicting each other. Moreover, the amount and type of resources available in the edge network could be very heterogeneous in terms of e.g., computational power, energy consumption, communication reliability, maximum bandwidth, number and type of devices. To this end, an orchestration of distributed AI/ML engines is crucial to find the most suitable Edge DNN architectural solution and deployment according to the high-level requirements to be met.

In detail, the orchestrator would need to allocate the AI models and tasks where considered more appropriate. This deployment surely benefits the latency, privacy, and energy efficiency requirements, because the huge amount of data is only processed and isolated at the edge locations saving a considerable amount of bandwidth and, thus, energy by means of not sending and, therefore, not storing (sensitive) data to the cloud. However, the AI model and task can suffer from a lack of performance and accuracy, since, in some cases, the edge server could not have the computational power and storage capacity as a centralised cloud server usually has. Nevertheless, the edge orchestration should not be limited to find the solution that best suits the requirements, but also a continuous monitoring of the performance of the distributed intelligence at edge network is very recommendable. The huge amount of data conveying to the edge devices could change their pattern over a long time and the AI models and tasks in some cases could underperform. As a result, the edge orchestrator could semi-autonomosly decide to either adjust or redeploy a more robust and reliable AI model at a specific edge location, increasing the performance of distributed intelligence.

The orchestration solution for distributed edge AI can be intended to be applicable to different 6G use cases that require the configuration and deployment of multiple AI agents, satisfying the different high-level requirements such as energy consumption, communication reliability and potentially others as well.

4.1.6 Compute-as-a-Service providing trustworthy and sustainable AI-based workload assignment

The aim of this section is to highlight the potential in-network AI functionality will have to support decision making towards discovering and selecting the network node(s) (either a device or a network infrastructure entity such as an edge cloud server) capable of executing a generated processing workload in a high-performance, trustworthy and energy-efficient manner. Such a workload may refer to an ML model learning task or to any generic user application involving a processing task, e.g., for storage and processing of large sensory data in an industrial environment aiming to formulate a DT of a factory floor, or in scenarios of immersive telepresence for enhanced interactions among humans, also possibly involving machines.

The topic of task offloading has been extensively covered by both academic literature and relevant standards specifications. Latest works involve the design of AI/ML-based solutions for workload offloading. The concept of CaaS is detailed in [D1.2]. In addition, [ZZL+19] focuses on a vehicular communication setup involving MEC infrastructure and proposes a deep Q-learning approach for designing optimal offloading schemes, jointly considering selection of target server

and determination of data transmission mode. In terms of performance requirements, the proposed solution takes into account the level of delay tolerance of the involved application. With regards to the need for reliable workload execution, [ZZL+19] proposes a redundant offloading algorithm to improve task offloading reliability in the case of vehicular data transmission failure. Nevertheless, the proposed design does not consider any network sustainability or network node trustworthiness criterion.

Further, the authors in [HFZ+19] focus on a MEC-enabled network deployment and propose a deep Q-network-based task offloading and resource allocation algorithm. Specifically, they consider that each mobile terminal has multiple tasks offloaded to an edge server and design a joint task offloading decision and bandwidth allocation optimisation problem to minimise the overall offloading cost in terms of energy, computation and delay cost. Although the principle of sustainability is acknowledged, the considered energy consumption model only involves radio energy consumption; on top of that, only a single cloud server is involved in system setup. Additionally, the authors in [LMZ+20], aiming to reduce task offloading delay, queuing delay, and handover cost in a vehicular environment, while simultaneously ensuring privacy, fairness and security, develop a scheme exploiting blockchain and smart contracts to facilitate fair task offloading and mitigate various security attacks and, on top of it, propose an online learning algorithm able to learn the long-term optimal strategy. This work, although aims to address the issue of trustworthiness for task offloading completely disregards the energy efficiency of the proposed scheme.

In terms of related standards, the ETSI Industry Specification Group (ISG) MEC has developed several use cases in Group Specification MEC 002 [EGM18], some of which focus on the topic of task offloading. For instance, there is a use case on multi-Radio Access Technology (RAT) application computation offloading; according to its description, the MEC system could help the application to select the most power-efficient RAT for the user device to improve the user experience in the network with multi-RAT coverage, apart from considering other performance indicators (e.g. offloading latency). Additionally, another use case is on in-vehicle MEC hosts supporting automotive workloads. One possible functionality of in-car MEC hosts deals with offloading processing-demanding tasks from vehicles to the network, e.g., relevant to computation-hungry applications such as Augmented Reality (AR), Virtual Reality (VR), AI and others. Nevertheless, [EGM18] does not touch upon AI/ML-assisted task offloading.

Future directions consist in the application of data-centric techniques to assist with the decision on whether to delegate a given workload to a different network entity (or distribute it across network entities) or address the workload locally at the network node it has been generated at. To pre-evaluate the energy efficiency of the potential workload addressment options across the network, an essential element of the solution will be to approximate the distribution of power consumption of the various compute nodes in the network over time for different workload generation patterns. Such knowledge can then be used to identify the network node that can host the workload with minimum energy consumption (sustainable operation). Of course, overall energy consumption needs to be considered, i.e., both the energy consumption of data transmission and the one due to workload processing.

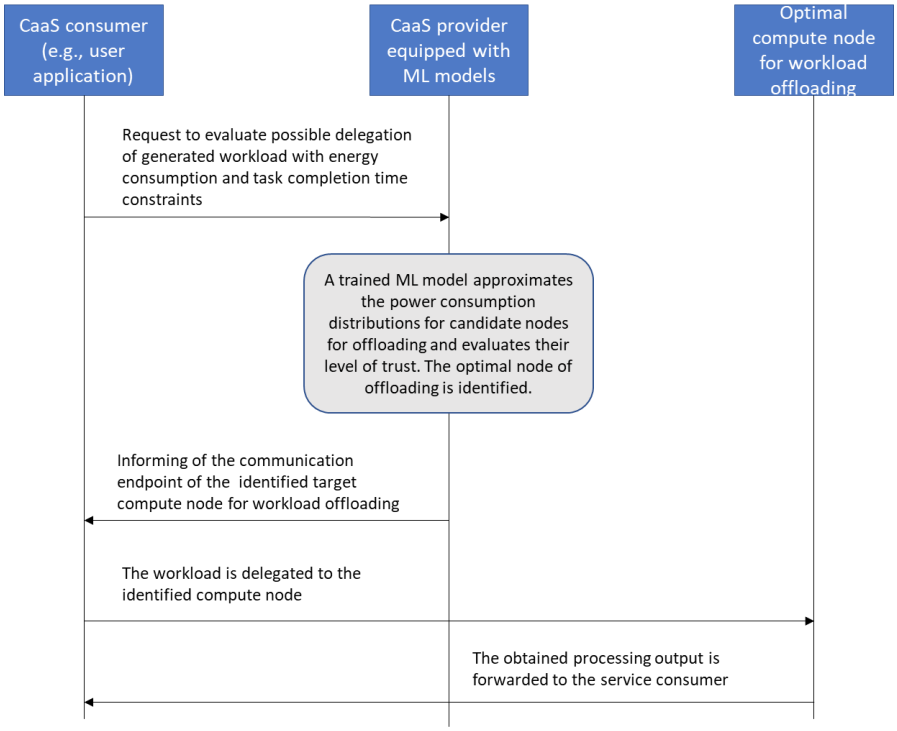

Figure 4-4: Basic principle on using CaaS for trustworthy and sustainable AI-based workload assignment

Similarly to per network node energy consumption distribution identification, AI/ML-based techniques can be applied to predict: (i) the processing load of a network node, as part of its compute, memory or storage capacity, which reflects the availability of the node to undertake a certain additional workload and reliability of output generation on time and (ii) whether the considered network node constitutes a trusted execution environment, that is, evaluating its capability to withstand security attacks that may evolve in time. The principle of the solution direction is shown in Figure 4-4.

The concept of AI-assisted CaaS is envisioned to be applicable to several 6G use cases, as the ones documented in section 2.6.2. For example, it can be part of a system design aiming to satisfy requirements of the "Interacting and cooperative mobile robots" use case due to expected incurred latency and reliability benefits. Relevant KPIs to be considered by new designs are the ones of end-to-end latency, AI agent availability and reliability, as well as network energy efficiency and in terms of KVIs, trustworthiness of the execution environment needs to be ensured before workload migration, as well as fairness of the solution.

4.1.7 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. Therefore, the investigation of the optimal resource utilisation and efficient knowledge sharing mechanisms are essential for the operation of AI functionality in wireless communication systems. The unprecedentedly extensive usage of wireless networks and the corresponding amount of data aggregated during their operation consist both a driving force and an enabler for data-driven methods, including AI/ML.

To minimise communication cost and maintain data privacy, attention has been concentrated to on the FL paradigm that consists of a local model updating with private datasets step, followed

by model averaging in a central server and broadcasting the consolidated model. However, this significantly increases the resource requirements for the central server and also reduces resilience to the case of a server malfunction. On the other hand, given the heterogeneity in the network, Decentralized FL (DFL) approaches suffer from slow convergence and degraded learning performance. A potential solution is presented in [LLV20], where DFL is performed via mutual knowledge transfer among the nodes. Another issue regarding the resource management in FL is the dependence on the constrained computational capabilities of edge devices. This impedes the training of complex models that offer increased accuracy. A group knowledge transfer training algorithm is proposed by [HAA20] generating promising results.

Blockchain has been gaining support in the last years with many applications. Lately, some approaches that combine FL with blockchain have been gaining ground. This way, privacy aware ML models at the core of a FL ecosystem can enable the entire network to learn from data in a decentralized manner [Nag19, CLZ20].

In order to ensure resource efficiency when deploying AI/ML methods in large scale wireless networks and, in parallel respect the distributed data pattern and data privacy requirement, different functionality allocations and knowledge sharing methods should be evaluated. Other essential factors to be considered include communication efficiency, learning quality, as well as resilience.

The application such methods facilitate multiple use cases discussed in section 2.6.2. The focus is mainly on the Immersive smart city, due to the large amount of data and devices involved, and on the Interacting and cooperative mobile robots, which require resource efficient data and model sharing. The KPIs that will depict the performance are primarily network energy efficiency and inferencing accuracy and the KVIs in scope are fairness and model complexity.

4.1.8 Semantic and goal-oriented communication approach for AI/ML at the edge

One of the goals of future 6G networks and services is to achieve system effectiveness and sustainability for AI/ML native applications. As pointed out in Section 4.1.1, an efficient training/inference of ML models and communication among AI agents 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 [MER21].

Indeed, the key challenge of previous mobile communication systems has been the correct reception of exchanged data, while targeting higher capacity, reliability and lower latency. Instead, with beyond 5G systems, the communication network is evolving towards a system capable of dealing with complex communicate-compute-control tasks, thus involving generation, transfer, and computation of large amounts of data, e.g., in the context of smart cities or AIaaS [D1.2] and, consequently, new requirements in terms of latency, communication bandwidth, etc. This data-to-intelligence loop is laying the foundations for the birth of connected intelligent machines – identified as a new driver for 6G. Today, while deploying 5G systems and kicking-off research on beyond 5G and 6G future networks [CBG+19], the need for a paradigm shift to goal-oriented communications begins to take shape to tackle, at some point, the bottleneck represented by the scarcity of resources, like spectrum and energy.

Semantic and goal-oriented communications go beyond the common paradigm of guaranteeing the correct transmission and reception of data, irrespective of the meaning they convey. The idea is that, whenever communication occurs to convey meaning or to accomplish a goal, what really matters is the impact that the received bits have on the interpretation of the meaning intended by

the transmitter or on the accomplishment of a common goal (e.g., a correct classification) [MER21], an aspect also tackled in WP7 [D7.1] with a focus on dependability and communication-control co-design. As an example, [SKA19] studies an edge learning system, with a Stochastic Gradient Descent (SGD) based algorithm running in an edge processor, to explore the trade-off between latency and accuracy, by optimising the packet payload size. In [MER21], novel ways are exploited to explore the trade-off between network energy consumption, delay and accuracy for different learning and inference algorithms. In this context, the exploration of how semantic learning can help 6G networks to improve their efficiency and sustainability is a research direction worth investigating further.

Previous research on semantic communications has focused on the potential offered by the introduction of a semantic level in the communication chain. In particular, several works propose alternative approaches, each aimed at emphasising different perspectives: philosophy of information [Flo02], logic and information [Dev95], information algebra [Koh12], information flow [BS97], quantum information theory [NC02], algorithmic information theory [Cal13] [Cha77]. In [WK05], the authors propose a method to perform semantic lossless data compression, as a way to produce a significant compression with respect to entropy-based encoders. Semantic data compression and the capacity of a semantic channel is studied in [BBD+11]. An end-to-end semantic communication framework incorporating semantic inference and physical layer communication problems is proposed and analysed in [GYS18]. A further extension is given in the recent work [XQL+20], where the authors, building on recent Natural Language Processing (NLP) tools, use a DNN to jointly learn a semantic/channel encoder, considering similarities between sentences. However, such approaches do not tackle the potential of semantic communication for non-natural languages and intelligences. The work in [CB21] identifies and justifies the need to explore the potential of semantic communications for future 6G networks and services. Very recently, the first potential benefits of semantic and goal-oriented communications for future 6G networks have been evaluated [CBG+19]. This preliminary work that started in 2021, has driven to the identification of a new possible class of 6G services: the Semantic Services [CBG+19], “semantic services will support all applications involving a share of knowledge between the interacting parties […], bringing a radical paradigm shift that will revolutionize wireless services from connected things to connected intelligences”. In Hexa-X project, novel solutions to achieve system effectiveness in the context of edge AI/ML will be investigated, possibly going beyond NLP, by exploring interactions among AI agents.

Two directions are worth to be addressed: a) The use of AI/ML to explore the semantic and goal-oriented approach aimed at enabling efficient communications among agents; b) Network optimisation, including communication and computation resources, aimed to enable training and inference with reduced communication resources, within low end-to-end delays, possibly with target performance. The use of the goal-oriented communication approach can help exploring typical trade-offs arising in this context. In this direction, among the expected benefits, a target will be to evaluate the positive impact of semantic communications on energy efficiency and/or uplink traffic reduction [OC16]. Several 6G use cases can be enabled, especially the ones involving interacting intelligent agents such as the "Interacting and cooperative mobile robots" use case. Also, applicability to the use case of "Flexible manufacturing" is of interest. From a KPI/KVI perspective, network energy consumption, latency, and inferencing accuracy represent the most relevant examples.

4.2 Enablers for in-network AI privacy, security and trust

As AI is becoming a required capability for both business and technology, trustworthiness is an inevitable requirement that needs to be addressed to avoid potential risks. Trustworthy AI ensures

the compliance with several requirements relevant to key principles [EC19]: human agency and oversight, technical robustness and safety (including security aspects), privacy and data governance, transparency (including explainability aspects), diversity, non-discrimination and fairness, societal and environmental wellbeing, and accountability. This compliance should be provided throughout the lifecycle of design, development, deployment, and operation. In this section we will provide more detail from a privacy, security and explainability point of view, on top of such considerations, as documented in [D1.2].

4.2.1 Privacy enhancing technologies for collaborative AI/ML

Collaborative AI/ML methodologies are a big step towards enabling privacy aware model training, as they aim joint model training by only sharing the necessary parameters and not client data. On the other hand, there are recent advances also from the attack sides, which result in more sophisticated privacy attacks, such as membership inference [SSS+17], property inference [MSD+19], and model extraction [TZJ+16] that can still be launched.

The adversarial goal of privacy attacks is to gain more information about the sensitive data, such as the training data and the ML model parameters. Privacy attacks can arise both at training or inference phase of ML and both server and clients can exploit the vulnerabilities. A malicious server may target to infer sensitive information from client updates, tamper with the training process or control the view of the clients on the global model. The aim of a malicious client could be to infer sensitive information of other clients benefiting from the inference output or to poison the data [WSR+20] or the model [FYB18].

Fortunately, Privacy Enhancing Technologies (PETs) providing a set of building blocks, namely differential privacy [DR14], homomorphic encryption [RAD78], secure multi-party computation [Yao82], and confidential computing using Trusted Execution Environments (TEEs) can be used as a remedy in such cases either in standalone or hybrid fashion.

Differential Privacy (DP) enables quantifying privacy by bringing a bound on the probability that two datasets can be distinguished. Using DP in the context of AI/ML leverages the privacy of individuals whose information is in a dataset. Record level or client level privacy is protected against attacks targeting to identify if individuals' data is being used during training. Therefore, DP is rigorously addressing the so-called membership inference attacks (i.e., given a data record and block-box access to a model, determine if the record was in the model's training dataset).

Homomorphic Encryption (HE) and Secure Multi Party Computation (SMPC) enable secure computation on the data between different parties without revealing sensitive data. HE is a form of encryption that allows the computation on ciphertext using specific operations without accessing to secret key nor requiring any decryption. The resulting computation persists in encrypted form until the keyholder decrypts the result. SMPC uses a set of cryptographic approaches such as Yao's Garbled circuits, oblivious transfer, and secret sharing to accomplish a joint computation over a function between parties using their sensitive inputs without revealing any information about their inputs to other parties. In the collaborative AI/ML context, these techniques provide secure construction of a global model; for example, in FL, the server can compute the aggregated global update, while not learning any information about clients' updates, with the help of SMPC.

Confidential computing technologies ensure that the data in use are protected against threats from malicious insiders with administrative privilege, direct access to hardware and, malwares that exploit bugs in the environment in which application runs on. This assurance is enabled by the use of advances of a TEE to build enclaves or trusted zones creating hardware-isolated and protected memory regions of the code and data. Privacy preserving solutions using TEEs require hardware capabilities such as memory isolation and memory encryption. Current technologies

like Intel SGX, ARM TrustZone, RISC-V or Sanctum provide these capabilities. In the collaborative AI/ML context, confidential computing can be used in different settings to protect deployed models from untrusted aggregators such as ML-as-a-service and the server in FL settings.

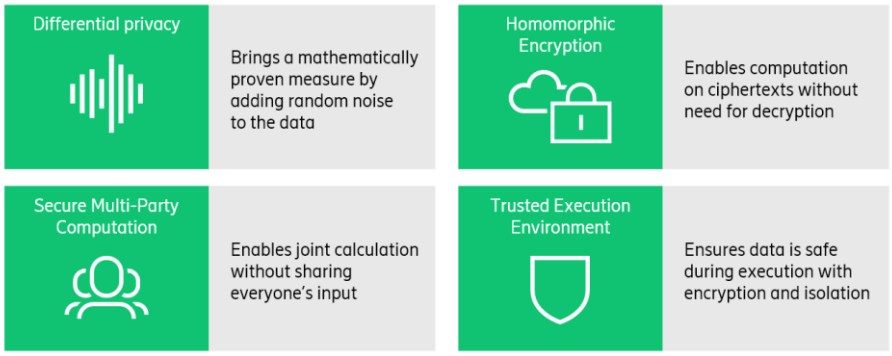

Figure 4-5: PETs at a glance

ML/AI for privacy as well as privacy for ML/AI are both regarded as important aspects in this direction and are anticipated as top strategic technology trends in 2021[3], since data protection is requested by both society and regulators. There are numerous published studies in both academia and industry to address privacy concerns both in theory and practice. Recently, application libraries, frameworks, and platforms on different privacy technologies started to be developed so that researchers can use to develop solutions to address privacy issues. [BLW08], [RWT+18], [ZDC+21], [ACP], [CGG+19] can be given as examples for such libraries and frameworks. A multitude of technology firms and governmental agencies invest heavily in enhancement on privacy enhancing technologies, e.g., DARPA recently granted $14.5M to develop privacy-preserving hardware accelerators[4].

Although there are considerable efforts to reflect advances, there is no single solution that covers all types of mitigations due to following reasons: (a) Each privacy solution addresses different types of threat models; (b) each collaboration setting has different trust relations; (c) different use cases lead to different requirements and specific constraints to consider, e.g., communication and computation capabilities of the clients might not be the same for each use case. Hence, as "one-size does not fit all", a better approach would be to first defining the requirements of the use case based on the threat model and incurred trade-offs and then select the appropriate PETs. In some cases, constructing a hybrid solution that uses multiple PETs instead of adopting only one of them may serve a better option.

Current privacy solutions are mostly proprietary proposals which makes the integration of different solutions together difficult. Also, utilisation of these solutions by non-expert users may be difficult. Thus, specifications and guidelines prepared by standardisation bodies would be very helpful to increase the deployment of such privacy enhancing technologies. Considering the significant interest in this area, standardisation bodies NIST and ETSI have started some activities on DP [NPE], Confidential Computing [CCC], and Privacy-Enhancing Cryptography [NPC].

[3] https://www.gartner.com/smarterwithgartner/gartner-top-strategic-technology-trends-for-2021

[4] https://www.prnewswire.com/il/news-releases/duality-technologies-awarded-14-5m-darpa-contract-to-develop-worlds-fastest-privacy-preserving-hardware-accelerator-301221040.html

4.2.2 Withstanding adversarial and poisoning attacks in network AI

The approach for AI in B5G/6G networks is heading towards a decentralised learning process. Methods for the collaborative training of DNN models have been proposed. For instance, [LLH+20] suggests a first phase of local model training and then, a second phase for the finalisation of the training through global aggregation of the updated parameters. In the last years, Google has introduced solutions that support this approach, namely TensorFlow Federated (TFF). TFF is an open-source framework for experimenting with ML and other computations on decentralised data [TFF21]. Nevertheless, it largely remains an open problem that certain generative ML models can craft systematic (evasion and poisoning) attacks against ML classifiers.

Data poisoning is a process by which an adversary injects malicious points in the training dataset to influence the learning process and degrade the algorithm's performance. There can be generation of adversarial training sequences that will degrade the classifier's accuracy. A sample adversary that can create a maliciously trained network is presented in [GLD+19]. Moreover, as described in [Goo16], there can be deep generative models that learn via the maximum likelihood principle, by constructing an explicit probability density distribution or by providing some way of "indirect interaction" with this probability distribution. In the FL case, the effect of poisoning can spread to various areas of the network, in which the model is distributed. The mitigation of this effect is largely an open problem, nevertheless, there are specific directions for future work.

To prevent such adversarial attacks, the use of Generative Adversarial Networks (GANs) is the cutting-edge approach, as described in [YU19]. According to the GAN paradigm, two major approaches have been applied in these cybersecurity studies: (a) the GAN is used to improve generalisation to unforeseen adversarial attacks, by generating novel samples that resemble adversarial data and can then serve as training data; (b) the GAN is trained on data with the goal of generating realistic adversarial data that can, thus, mislead a security system.

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. In the envisioned approach, data sanitisation could rely on techniques such as anomaly detection, as well as on more complex methods, like removal of negative impact (elimination from training of those points that have a substantial negative impact on the classification accuracy) [BNJ+10], training with micromodels (so as to reduce the risk of attack), use of GANs. Leveraging on the GAN paradigm, realistic samples that resemble adversarial data can be generated and be used in the training, for enabling FL models to obtain a sense of immunity to the poisonous attack.

The issue under investigation is mostly applicable in the use case of Immersive smart city quantity of user devices, considering the quantity of user devices. Explainability and fairness are the most relevant KVIs, while AI agent reliability and inferencing accuracy the respective KPIs.

4.2.3 Explainable AI for mitigation of biased decisions

Explainable AI (XAI) is the process of explaining why an AI agent performs, after an internal processing, a certain decision to the final user in understandable terms. In [DBH18], two methods for explainability are described: transparency-based and post-hoc.

Transparency is one of the fundamental properties that enables interpretability and, thus, the explanation of the decision-making process of AI system. However, in some cases, it could be too complex for describing to the final users in understandable form. In the post-hoc mechanism, the information is extracted after the decision process has taken place. The AI system is treated

like a black box and for each decision the output is analysed trying to identify the rationale behind it. A general explanation for the decisions is given by the importance of involved data features and, in particular, either NN layers or neurons have a crucial role in the decision. For instance, the work in [WTK17] analyses why a particular image has correctly been classified as a rooster. The answer relies on the importance of the pixels analysed by the AI agent, in particular, those forming the red comb and the wattle.

On the other hand, it is well-known that AI systems are not perfect in the decision-making process as they may be sometimes unfair, as biased decisions could be taken, because they could tend to inherit biases that are already present in the training data set. In some cases, these decisions could become even harmful in domains such as health care, assurance, self-driving, and others. To address this problem, the EU has defined the Ethics Guidelines for Trustworthy AI [EC19] related to the technical robustness, safety, diversity, non-discrimination and fairness, of an AI system. The main idea behind the guidelines is to design and build an AI system to prevent unacceptable harm and avoid discrimination introduced by a possible biased dataset. Thus, to overcome the biased decision problem, actions could be applied to the dataset.

A straightforward solution to overcome the biased decision problem, once the origin of the issue has been identified, is the pre-processing of the dataset in order to obtain an unbiased trained model. However, due to the lack of uniform data or the impossibility to remove the sensitive data that introduces bias, it would not be always possible to modify the original dataset. For this reason, other solutions have been investigated oriented to the control of the chain composing the AI decisions or to alter the model parameters. For instance, [Ch19] proposes a path-specific counterfactual method, where AI systems perform a correction on the variables that are descendants of sensitive variables through unfair decision paths.

In the 5G context, the AI techniques are used in several areas of network control and management, e.g., to optimise the resource utilisation of the network infrastructure, for virtual function placement, and for traffic and mobility management or network slice management. AI-based data-driven solutions are expected to be deployed at the 5G Core Network, in the NWDAF, or at the management system in the Management Data Analytics Function (MDAF). However, the applicability of explainable AI techniques for biased decisions mitigation in such functions is still to be investigated, e.g., in terms of internal functional decomposition and interfaces. This would be critical to guarantee the fairness of the network management and orchestration, without biased decisions that may lead to favouring certain categories of traffic or network slices.

At the state of the art, the NWDAF defined in 3GPP specifications (Rel. 17) for collecting and analysing data from other 5G Core Network functions (NFs) is functionally split in two components. The Analytics Logical Function (AnLF) performs inference, derives analytics information and exposes analytics service, while the Model Training Logical Function (MTLF) is dedicated to training ML models, which can be then consumed on-demand by the AnLF. A similar decomposition can be easily applied also to the management system at the MDAF level.

In this scenario, the source dataset used for training the model at the MTLF could be biased and this could result in unfair decisions taken at the AnLF. However, the adoption of XAI algorithms and techniques could help in understanding the decisions' motivation, detecting potential unfairness and identifying its origin. Such information can then be used for taking actions to actively remove the unfairness, e.g., applying a self-adjustment mechanism to mitigate the biased "knowledge" the ML models can have. In this direction, the current NWDAF (or MDAF) architecture can be extended as depicted in Figure 4-6. Two additional functions are introduced to complement the AnLF and MTLF actions and to realise an automated closed-loop for detecting unfairness and its origin (the *Continuous verification* block) and trying to mitigate the issue (the *Unfairness Mitigator* block) acting at the model training level, or through a post processing of the

analytics decisions. Such closed-loop approach could be fully automated, with a human-in-the loop intervention with the purpose of make the continuous verification and unfairness mitigator blocks more robust.

To conclude, explainability and fairness are two most applicable KVIs to take into consideration for 6G systems, with the final aim to mitigate the bias decision problem of MTLF and AnLF.

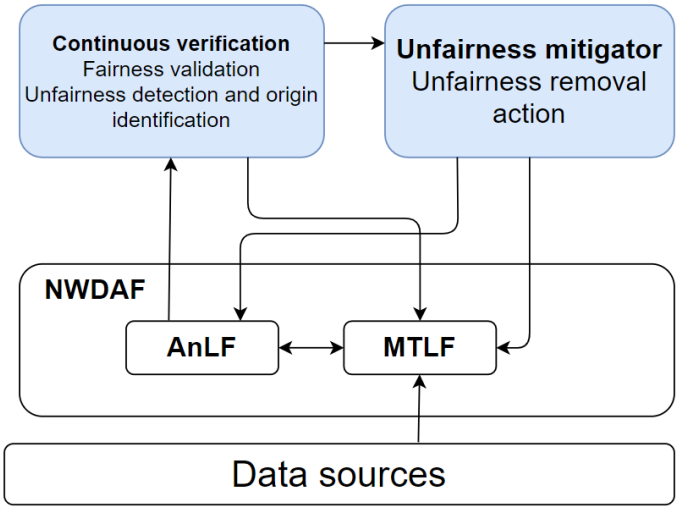

Figure 4-6: NWDAF extension to mitigate biased decisions

4.2.4 Fed-XAI: Federated explainable artificial intelligence

Since early works in FL literature [MMR+17] [KMR+16], most solutions revolve around the original proposal of federated averaging (FedAvg), as a protocol for executing SGD in a federated manner. Specifically, [MMR+17] showed that DNN models can be collaboratively trained for tackling image classification and language modelling tasks. While DNNs have achieved unprecedented levels of performance in various application domains, they are generally considered opaque or black-box models due to their huge number of parameters and non-linear modelling: as such, they do not feature inherent explainability. Authors in [ADD+20] provide a detailed overview regarding the two strategies for achieving explainability, namely the design of inherently interpretable models (i.e. "transparent box design" strategy) and the adoption of post-hoc explainability techniques (i.e. the "explaining black-box" strategy). The following popular ML models are generally considered transparent, i.e., understandable for a human: linear/ logistic regression, decision trees, k-nearest neighbours, rule-based learners, generalised additive models, and Bayesian models. In fact, they feature at least one of the following three properties [ADD+20]: algorithmic transparency, decomposability and simulatability. ML models that do not meet any of the requirements imposed to be defined as "transparent" require the use of some post-hoc techniques to explain their decisions, e.g., feature relevance, local explanation or visualisation. Some shallow models, e.g., tree ensembles, random forests, SVMs, and deep models, e.g., DNNs, CNNs, Recurrent NNs (RNNs), belong to this category.

Finally, it should be underlined that the performance of a model and its transparency are typically conflicting objectives and, thus, accuracy-oriented solutions are often deemed as hard to interpret. This trade-off is shown in Figure 4-7, with respect to some popular ML models.

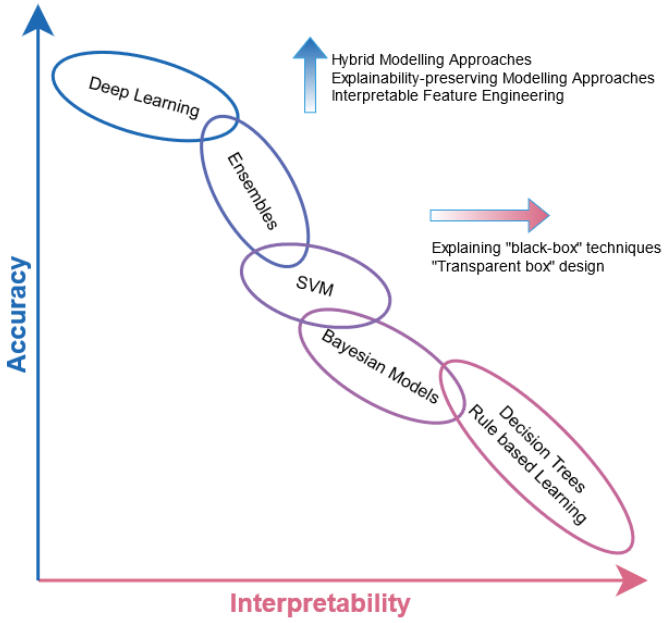

Figure 4-7: Trade-off between model interpretability and performance. Inspired by [ADD+20]

It is worth noting that decision trees and rule-based systems lie in the region of "high interpretability". When these models handle linguistic variables (i.e., the input variables are categorical or undergo a discretisation step), interpretability is further enhanced, allowing a direct human interaction thanks to the explicit semantic understandability. In this context, linguistic fuzzy models provide a natural linguistic representation of numeric variables and typically outperform their crisp counterparts in scenarios with some degree of noise and/or uncertainty.

Some recent works [SLC+20], [Guo20] have reviewed the current status of AI-enabled cellular networks. Specifically, authors in [SLC+20] provide an overview of the key thrusts in AI for the PHY, MAC and network layers, with applications ranging from channel estimation and prediction, to dynamic spectrum access and resource management and scheduling. Interestingly, the overview of the authors highlights that, in the last few years, the most prominent tools adopted in this context rely on models based on NNs, e.g., DNN, CNN, RNN, DRL. A similar overview is presented in [Guo20], with a specific focus on the trustworthiness of AI/ML techniques. This survey, like the previous one, cannot be considered exhaustive within the vast literature on the subject. Nevertheless, it reports some representative examples of the adoption of AI methods in wireless communication along with the degree of explainability that is generally accorded to the methods themselves. Out of a total of 14 papers analysed, 11 of them involve methods with a degree of explainability that is defined as none, very low or low.

Despite few exceptions, the reviews discussed in the previous paragraphs reveal that much of the core methods involving AI in wireless network setting consist of accuracy-oriented solutions that do not provide inherent explainability.

Future investigations stem from the following considerations: AI/ML-based solutions are expected to be an essential component of B5G/6G technology. Actually, they have also been extensively exploited in the last decade to address several problems in the field of wireless networks. Nevertheless, existing approaches ignore one or both of the following requirements: a) the preservation of privacy of data owners, while collaboratively training ML models and b) the

explainability of the models. To this aim, the Fed-XAI vision is about devising methods and approaches compliant, at the same time, with FL and explainable AI paradigms.

A possible direction to be investigated could deal with the federated implementation of inherently explainable (transparent) algorithms. Linguistic models may prove appropriate to ensure an adequate level of interpretability; tools from fuzzy set theory may also improve performance in contexts characterised by noise and uncertainty. Indeed, the challenges of the FL setting will need to be addressed: data from different sources may have different distributions (non i.i.d.) and volumes; their number can grow fast and their participation in the federation may be unstable.

Notably, Fed-XAI can be regarded as an enabler for several families of use cases envisioned for 6G. Inferencing accuracy represents the most relevant KPI, which must be pursued together with the KVI of explainability, deemed as a crucial requirement for trustworthiness. Model complexity can be considered as a proxy for the interpretability level and may be associated with other XAI metrics (e.g., based on surveys) to evaluate explainability.

4.3 AI powered network operation

In this section, the aim is to concentrate on the benefits of widely enabling AI/ML functionality to 6G networks by design. In section 4.3.1, the topic of AI-based management and orchestration for behaviour-driven adaptation is elaborated, targeting automated network operation, while, in section 4.3.2, the goal is to present the advantages of AI/ML technology when used for security enhancements in a 6G network, focusing on intrusion detection.

4.3.1 AI-based management and orchestration for behaviour-driven adaptation

One of the main aspects considered in the scope of the H2020 ICT-52 call is the long-term transformation of networks into a distributed smart connectivity platform with high integration with (edge) computing and storage resources. Focusing on such platforms, investigation should lead to solutions where processes and applications are dynamically supported depending on the information flows and application requirements. The Hexa-X approach to this regarding the network management and orchestration is to provide a framework for dynamically supporting reliability and resilience in changing requirements providing the so-called “continuum” device-edge-cloud management to address mobility and resource utilisation. Particularly, one of the main objectives in WP6 is to demonstrate algorithms for data-driven device-edge-cloud continuum management and, also provide implementations of continuum management of device, edge, RAN, and cloud as one of the main measurable results.

In practice, this concept of "continuity" considering cloud, edge and devices introduces a relevant innovation regarding management and orchestration: now it is necessary to consider not only core and edge network resources (as it happens 5G), but also what is typically called the “extreme edge” [PML+19], i.e., all those end-user devices beyond the access network[5]. One of the main challenges associated to this is the need to deal with an increasing number and diversity of end-devices [PEE21], which is also associated to the development of the IoT technologies.

In this context, AI/ML techniques are considered a key enabler due to the increased complexity and heterogeneity of data sources in this new extreme edge domain, and because of the benefits

[5] This “extreme edge” is also sometimes referred to as “far edge” [LEM20] or “deep edge” [WG20] [ERI19]

offered by these technologies for dealing with large amounts of data that can be generated from heterogeneous sources and with different formats [YCX+20]. Where the regular algorithms development process cannot be suitable for processing large amount of data, AI/ML techniques are seen as an appropriate technique to deal with this [EAA20].

AI/ML techniques are already gaining attention regarding network management and orchestration (MANO) [D6.1]. An example is the Industry Specification Group (ISG) on Experiential Networked Intelligence (ENI) [ENI21] launched by ETSI in 2017 with the main purpose of defining an architecture based on AI/ML techniques, context-aware and metadata-driven policies according to the "Observe-Orient-Decide-Act" (OODA) control loop model [FGC12] [EGE19]. Different works have been performed based on this model; for instance [WFC18] offers an overview about requirements and different use cases; [GGD+19] proposes a programmable architecture based on network slicing and [BGP+20] defines a framework with specific AI-based algorithms serving for different purposes.

Under the ETSI umbrella it can also be found some progress in this direction on the Open Source MANO (OSM) orchestration platform [OSM20], specifically regarding the integration of AI agents as part of the platform itself. Those AI agents would be attached to the deployed NFs by specific NF execution environments. A PoC has been conducted in order to prove the potential usage of this functionality [OSM21]. In a different direction current Release TEN [OSMRT20] also fulfils ETSI's Zero touch network & Service Management (ZSM) [ZSM21] which is more focused on AI-supported automation; in this case, the target is to reach a full E2E automation of network and service management, which will bring new requirements for E2E architectures designed for AI algorithms, closed-loop automation and optimised data-driven ML. ZSM management functions support data-driven automation based on closed-loop and integration of AI/ML techniques.

Another interesting approach in this direction is the one by the Open Network Automation Platform (ONAP) [ONA21], which is also a platform for automation, orchestration and management providing a modular architecture. ONAP is at an early stage regarding the application of AI/ML techniques although initial steps are being taken to make the orchestrator able to trigger automated actions when associated AI/ML modules detects certain threats regarding latency or security [OAIDL18].

Beyond specific approaches AI/ML techniques in general can improve management and orchestration. For instance, using the main learning approaches in AI/ML, the following issues could be addressed:

- Supervised algorithms [KOT07], to trigger management and orchestration actions based on time series forecasting or complex pattern recognition techniques;
- Unsupervised learning [GBL+03], to clustering data or threat identification;
- Reinforcement learning [SB17], to implement automated control loops;
- FL [NDR20], to deploy distributed agents on the network infrastructure and by using distributed training data sets.

Also, 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. This predictive orchestration can rely on identifying known risk factors based on historical events using AI/ML models that can also improve their predictive capabilities over time, making stronger correlations between the risk factors that have demonstrated a proper predictive potential.

But beyond proactive scaling actions AI/ML techniques can be used also to enhance the customer experience and network optimisation in other different ways, e.g.:

- Closed-Loop Automation actions.
- Performing automated NF placement actions (e.g., to deploy NFs on the edge or core networks).
- Implementing proactive alerting strategies.
- Hidden patterns discovery (using unsupervised techniques) that could be used by the OSS/BSS teams to implement new operational or business strategies.
- Preventive incident analysis.

Some of the main challenges related to the usage of AI/ML in the scope of network management and orchestrations are:

- Automation, use of resources must be done efficiently by providing network dynamicity, reliability and resilience.
- Explainability, orchestration actions based on regular (legacy) programmatic approaches are typically easily explainable, i.e., if the service is well developed ultimately there is always a set of log files that can be used to explain why an algorithm has executed one action or another. However, AI/ML models are sometimes a sort of black-boxes correlating inputs and outputs by a set of mathematical operations without a clear semantic associated to them. This may make it difficult to provide explanations about certain orchestration actions (NF migration, placement or other) that the AI could automatically perform. To address such problem specific Explainable AI (XAI) techniques are available (see 4.2.3) and should be considered as part of the management and orchestration system.
- Monitoring, due the huge number of devices to monitor (e.g., considering the extreme edge devices) and due the convenience to combine both: infrastructure metrics (CPU, RAM, network traffic…) and the data plane metrics. This is perhaps one of the most challenging to being address, since current NFV orchestrators (e.g., OSM or ONAP) are typically limited to process a well-defined set of infrastructure metrics, instead offering a more open data model able to integrate user-defined metrics that could bring richer AI/ML models.
- Security, the management and orchestration functionality should provide liability to mitigate security risks.

4.3.2 AI for Network Security: intrusion detection system architecture and detection procedures

The general problem of anomaly detection and classification is a recurrent thematic for networks supervision. 5G systems propose a renewal with the integration of three technological components: 3GPP RAN, TCP/IP protocols, virtualisation frameworks. A whole 5G infrastructure inherit the cyber-risks from each component. We propose here an integrated vision of data collection and anomaly analysis to deal with this situation.

In this part, two problems are elaborated, i.e., (a) the probing and storing of network events and (b) data representation and decision techniques.

4.3.2.1 Probing and storing of network events

With regards to the problem 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.

If many attacks can be detected with a one-shot observation (Heartbleed), sophisticated Advanced Persistent Threats (APTs) require information correlations in time or space (Botnet, Distributed Denial of Service - DDoS, Domain Generator Algorithms, Poisoning, Spoofing), which implies both localised and distributed storage and processing. What is more, these operations are constrained by detection/classification in real time (multiple request analysis) and data base finite size.

In terms of prior work in the area, the National Institute of Standards and Technology (NIST) Framework [NCF18] yields the general security process involving an IDS, while the MITRE-Common Vulnerabilities and Exposures (CVE) data base [Mit21] updates the known vulnerabilities and the Exploit Database [ED21] proposes attacks for penetration testing activities. Atomic functionalities are available in open source, such as the following: protocol interpretation with Wireshark for packet captures (PCAP format), Forensic / real-time IDS protocol events analysis (Zeek, Suricata), Forensic Post-Attack analysis (Network Miner) based on attack modelling, and Data Base storage for collected network logs (MySQL, SQLite).

System design Security Information Event Management (SIEM) tools are also available, such as Local Rock Network Security Monitor [NSM] and based on distributed and collaborative IDS [CS21].

The identified gap is that state-of-the-artwork proposes only partial solutions to the problem setting. Hence, solutions are needed to propose and prototype an architecture merging these different characteristics. In addition, todays IDSs are not perfectly designed for detecting unknown attacks or even for classifying them (there are more non malicious anomalies than malicious ones).

4.3.2.2 Data representation and decision techniques

Considering 6G network cybersecurity, SIEM will have to face an increasing amount of data to analyse, which is expected to be several orders of magnitudes larger than in 5G. Beyond the inherent legal constraints in storage communication data, 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 tricky to perform on encrypted data.

Along with this limitation, security management must also handle data aggregation according to the temporal and spatial dimensions of the network. This is because DDoS detection requires correlating information from different nodes over a short period of time, while APTs can only be detected by correlating information from different time periods to a given node.

As a consequence, an essential question in cybersecurity can be formulated as how to summarise efficiently network data, based on packet captures, to both guarantee their storage capability and accurate real-time threat analysis. In this context, accuracy is understood as reliable false alarm and detection or classification probabilities, for a large variety of potential abnormal network behaviours.

In other words, the solution direction is to apply data representation techniques (AI concepts) to cybersecurity. As a methodological guideline, it is useful to assess a taxonomy with typical

network behaviour extracted and modelled from CVE data bases. A proposal for such a taxonomy could be:

- Behaviour detectable on one node in a short time scale (requiring low local memory and processing);
- Behaviour detectable on one node on a long-time scale (requiring heavy local memory and processing);
- Behaviour detectable on several nodes, i.e., large spatial scale, and short time scale (requiring heavy and distributed memory and processing)

Network packet captures encompass miscellaneous type of data: bytes (bytes exchanged, packet error rates, delays, etc.) and categorical data (flag, protocol modes, etc.). It is then natural to envision modern techniques for data modelling and representation. They enable information compression, decorrelation and denoising with latent variables: variational auto-encoders [KW13], GANs in image/signal processing (numerical data) and embedding in NLP (categorical data) [BDV+03]. The Statistical Theory supplies rules for optimised decision in detection/classification problems.

The purpose is to build upon state-of-the-art solutions and expand to cyber cases involving both numerical and categorical data. As a strategy, time-evolutional compression in data bases should be considered to compensate short-term exhaustivity by redundancy in APT cases.

5 Conclusions and Recommendations

This report 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 fundamental AI concepts and a summary of common practices. The potential applications were investigated, starting with applications in the air interface and continuing with in-network learning methods.

Regarding AI-based air interface design, four main pillars were considered. The first one is novel, data-driven transceiver design approaches, focusing on hardware impairments in the transmitter and receiver RF chains. Then, AI-driven transmitters were considered with an investigation of AI optimisation of beamforming and RL methods for fast initial access. On the subject of AI-driven receivers, multiple channel estimation methods were discussed and channel decoding was explored, as well as receiver side processing as a single block. Finally, concerning AI-driven radio interface functionality, this document examined different approaches for RRM, cell-free and distributed massive MIMO systems, as well as model predictive control of antenna systems.

The topic of in-network learning is organised in three main parts. The subsection of joint communication and computation co-design investigated different approaches for distributed learning taking into account aspects of efficiency and resilience. Subsequently, enablers for in-network AI privacy, security and trust were analysed including privacy concerns, explainability features and resistance to attacks. Lastly, the topic of AI-based management and orchestration for behaviour-driven adaptation was elaborated along with the one of AI for network security focusing on the identification of network intrusions.

Summing up, the following consolidated recommendations are derived:

- 6G network architecture should enable and support “end-to-end learning”, i.e., learning and optimizing 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**.
- 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.

6 References

[22.874] 3GPP TR 22.874, " Study on traffic characteristics and performance requirements for AI/ML model transfer in 5GS {Release 18)", v2.0.0, Jun. 2021.

[23.501] 3GPP TS 23.501, "System architecture for the 5G System (5GS); Stage 2 (Release 17)", V17.1.1, Jun. 2021.

[23.748] 3GPP TR 23.748, "Study on enhancement of support for Edge Computing in 5G Core network (5GC) (Release 17)", v17.0.0, Dec. 2020.

[29.520] 3GPP TS 29.520, 5G System; Network Data Analytics Services; Stage 3 (Release 17), v17.1.0, Dec. 2020.

[AA19] M. Alrabeiah and A. Alkhateeb, “Deep learning for TDD and FDD massive mimo: Mapping channels in space and frequency,” in 2019 53rd Asilomar Conference on Signals, Systems, and Computers. IEEE, 2019, pp. 1465–1470.

[AA20] M. Alrabeiah and A. Alkhateeb, “Deep Learning for mmWave Beam and Blockage Prediction UsingSub6 GHz Channels,” IEEE Transactions on Communications, vol. 68, no. 9, pp. 5504–5518, 2020.

[ABG+19] E. Abebe, D. Behl, C. Govindarajan, Y. Hu, D. Karunamoorthy, P. Novotny, V. Pandit, V. Ramakrishna, and C. Vecchiola, "Enabling enterprise blockchain interoperability with trusted data transfer (industry track)," in Proceedings of the 20th International Middleware Conference Industrial Track, pp. 29-35. 2019.

[ACP] Galen Andrew, Steve Chien, and Nicolas Papernot. TensorFlow Privacy. https://github.com/tensorflow/privacy.

[ADD+20] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." Information Fusion 58 (2020): 82-115.

[AGST13] Aswani, A., Gonzalez, H., Sastry, S. S., & Tomlin, C. (2013). Provably safe and robust learning-based model predictive control. Automatica, 49(5)

[AH1813] F. A. Aoudia et J. Hoydis, « End-to-End Learning of Communications Systems Without a Channel Model », arXiv:1804.02276, 2018.

[AH19] F. A. Aoudia and J. Hoydis, "Model-free training of end-to-end communication systems," in IEEE Journal on Selected Areas in Communications, vol. 37, no. 11, pp. 2503-2516, Nov. 2019.

[AH21] F. A. Aoudia and J. Hoydis, “End-to-end learning for OFDM: From neural receivers to pilotless communication,” 2021. [Online]. Available: https://arxiv.org/abs/2009.05261

[Alp20] E. Alpaydin, “Introduction to machine learning”, MIT press, 2020.

[ASR20] S. Aliet al., “6G White Paper on Machine Learning in Wireless Communication Networks,” University of Oulu, 2020. [Online]. Available: http://arxiv.org/abs/2004.13875

[BA17] Stefano Buzzi, & Carmen D’Andrea, “Cell-free Massive MIMO: User-Centric Approach”, IEEE Wireless Communication Letters, August 2017.

[BA17] S. Buzzi and C. D'Andrea, "Cell free massive MIMO: User centric approach," IEEE Wireless Communications Letters, vol. 6, no. 6, pp. 706–709, 2017.

[BB12] Bernardini, Daniele, and Alberto Bemporad. "Energy-aware robust model predictive control based on noisy wireless sensors." Automatica 48.1 (2012): 36-44.

[BBD+11] J. Bao, P. Basu, M. Dean, C. Partridge, A. Swami, W. Leland, and J.A. Hendler. "Towards a theory of semantic communication." In 2011 IEEE Network Science Workshop, pp. 110-117. IEEE, 2011.

[BC20] W. Saad, M. Bennis, and M. Chen, "A vision of 6g wireless systems: Applications, trends, technologies, and open research problems," IEEE Network, vol. 34, no. 3, pp. 134–142, 2020.

[BDD+20] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012 , last accessed 16.06.2021

[BDF+21] S. Buzzi, C. D'Andrea, M. Fresia, Y. -P. Zhang and S. Feng, "Pilot Assignment in Cell-Free Massive MIMO Based on the Hungarian Algorithm," in IEEE Wireless Communications Letters, vol. 10, no. 1, pp. 34-37, Jan. 2021

[BDV+03] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. "A neural probabilistic language model." The journal of machine learning research 3 (2003): 1137-1155.

[BGL03] A Benczúr, U Glässer, T Lukovszki. Formal description of a distributed location service for mobile ad hoc networks. International Workshop on

[BGM+20] D. Belot, J.L. González Jiménez, E. Mercier, and J.B. Doré. "Spectrum Above 90 GHz for Wireless Connectivity: Opportunities and Challenges for 6G." Microwave Journal 63, no. 9 (2020).

[BGP+20] Bega, D., Gramaglia, M., Perez, R., Fiore, M., Banchs, A., & Costa-Pérez, X. (2020). AI-Based Autonomous Control, Management, and Orchestration in 5G: From Standards to Algorithms. IEEE Network, 34(6), 14–20. https://doi.org/10.1109/MNET.001.2000047, last accessed 23.04.2021

[BHP+20a] A. Buchberger, C. Häger, H. Pfister, L. Schmalen, and A. Graell I Amat, "Pruning neural belief propagation decoders," IEEE International Symposium on Information Theory, 2020.

[BHP+20b] A. Buchberger, C. Häger , H. Pfister, L. Schmalen, and A. Graell i Amat, "Learned decimation for neural belief propagation decoders," arXiv, 2011.02161, 2020.

[BKP19] A.A. Benczúr, L. Kocsis, R. Pálovics (2019) Online Machine Learning Algorithms over Data Streams. In: Sakr S., Zomaya A.Y. (eds) Encyclopedia of Big Data Technologies. Springer, Cham. https://doi.org/10.1007/978-3-319-77525-8_329

[BLW08] Bogdanov, D., S. Laur and J. Willemson. "Sharemind: A Framework for Fast Privacy-Preserving Computations." IACR Cryptol. ePrint Arch. 2008 (2008): 289.

[BNJ+10] M. Barreno, B. Nelson, A.D. Joseph, and J.D. Tygar, “The security of machine learning”, Machine Learning, Vol. 81, 2010, pp. 121–148, DOI 10.1007/s10994-010-5188-5.

[BRR+20] I. Be’Ery, N. Raviv, T. Raviv, and Y. Be’Ery, “Active deep decoding of linear codes,” IEEE Transactions on Communications, vol. 68, no. 2, pp. 728–736, 2020.

[BS20a] E. Björnson and L. Sanguinetti, "Making Cell-Free Massive MIMO Competitive With MMSE Processing and Centralized Implementation," in IEEE Transactions on Wireless Communications, vol. 19, no. 1, pp. 77-90, Jan. 2020

[BS20b] E. Björnson and L. Sanguinetti, "Scalable Cell-Free Massive MIMO Systems," in IEEE Transactions on Communications, vol. 68, no. 7, pp. 4247-4261, July 2020

[BS97] J. Barwise, and J. Seligman, “Information Flow: The Logic of Distributed Systems”, Cambridge University Press (1997).

[CA13] Camacho, Eduardo F., and Carlos Bordons Alba. Model predictive control. Springer science & business media, 2013.

[CAA20] G. Charan, M. Alrabeiah, and A. Alkhateeb, “Vision aided dynamic blockage prediction for 6G wireless communication networks,” 2020. Available: https://arxiv.org/abs/2006.09902

[Cal13] C.S. Calude, “Information and Randomness: An Algorithmic Perspective”, Springer Science & Business Media (2013).

[CAL94] Cohn, David, Les Atlas, and Richard Ladner. "Improving generalization with active learning." Machine learning 15.2 (1994): 201-221.

[CB21] E. Calvanese Strinati and S. Barbarossa, “6G networks: Beyond Shannon towards semantic and goal-oriented communications”, Computer Networks Journal, Vol 190. ISSN 1389-1286, DOI: https://doi.org/10.1016/j.comnet.2021.107930. February 2021.

[CBG+19] E. Calvanese Strinati, S. Barbarossa, J.L. Gonzalez-Jimenez, D. Ktenas, N. Cassiau, L. Maret, and C. Dehos. "6G: The next frontier: From holographic messaging to artificial intelligence using subterahertz and visible light communication." IEEE Vehicular Technology Magazine 14, no. 3 (2019): 42-50.

[CCC] Confidential Computing Consortium, https://confidentialcomputing.io/

[CGG+19] Chillotti, I., Gama, N., Georgieva, M., & Izabachène, M. (2019). TFHE: Fast Fully Homomorphic Encryption Over the Torus. Journal of Cryptology, 33, 34-91.

[CGH+17] S. Cammerer, T. Gruber, J. Hoydis, and S. ten Brink, “Scaling Deep Learning-based Decoding of Polar Codes via Partitioning,” ArXiv170206901, 2017.

[Ch19] Silvia Chiappa, Thomas P. S. Gillam. “Path-Specific Counterfactual Fairness". arXiv:1802.08139

[Cha77] G.J. Chaitin, “Algorithmic information theory”, IBM J. Res. Dev., 21 (1977), pp. 350-359.

[Chi19] S. Chiappa, "Path-specific counterfactual fairness." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 7801-7808. 2019.

[CHW+20] Y. Cui, X. Huang, D. Wu and H. Zheng, "Machine Learning based Resource Allocation Strategy for Network Slicing in Vehicular Networks," 2020 IEEE/CIC International Conference on Communications in China (ICCC), 2020, pp. 454-459.

[CLZ20] H. Chai, S. Leng, Y. Chen and K. Zhang, "A Hierarchical Blockchain-Enabled Federated Learning Algorithm for Knowledge Sharing in Internet of Vehicles," in IEEE Transactions on Intelligent Transportation Systems, doi: 10.1109/TITS.2020.3002712.

[CMG+149] K. Cho, B. van Merri¨enboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1724–1734, 2014.

[CMW+21] Yejian Chen; Jafar Mohammadi; Stefan Wesemann; Thorsten Wild; "Turbo-AI, Part II: Multi-Dimensional Iterative ML-Based Channel Estimation for B5G," accepted by 2021 IEEE 93rd Veh. Technol. Conf. (VTC'21 Spring), Helsinki, Finland, April 2021.

[CS21] CrowdSec, https://crowdsec.net

[CVK17] P. Casas, J. Vanerio and K. Fukuda, "GML learning, a generic machine learning model for network measurements analysis," 2017 13th International Conference on Network and Service Management (CNSM), 2017, pp. 1-9.

[CWL+19] Z. Chang, Y. Wang, H. Li, and Z. Wang, "Complex CNN-based equalization for communication signal," in 2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP), July 2019, pp. 513–517.

[D1.2] Hexa-X Deliverable D1.2, "Expanded 6G vision, use cases and societal values". Online: https://hexa-x.eu/wp-content/uploads/2021/05/Hexa-X_D1.2.pdf

[D6.1] Hexa-X Deliverable D6.1, "Gaps, features and enablers for B5G/6G service management and orchestration". Online: https://hexa-x.eu/wp-content/uploads/2021/06/Hexa-X_D6.1.pdf

[D7.1] Hexa-X Deliverable D7.1, "Gap analysis and technical work plan for special-purpose functionality". Online: https://hexa-x.eu/wp-content/uploads/2021/06/Hexa-X_D7.1.pdf

[DBH18] F. K. Došilović, M. Brčić and N. Hlupić, "Explainable artificial intelligence: A survey," 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2018, pp. 0210-0215, doi: 10.23919/MIPRO.2018.8400040.

[DCH+18] S. Dörner, S. Cammerer, J. Hoydis, and S. t. Brink, "Deep learning based communication over the air," IEEE Journal of Selected Topics in Signal Processing, vol. 12, no. 1, pp. 132–143, Feb 2018.

[Dev95] K. Devlin, "Logic of Information", Cambridge University Press, 1995.

[DKG19] M. Dias, A. Klautau, N. González Prelcic, and R. W. Heath, "Position and lidar aided mmwave beam selection using deep learning," in2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 2019, pp. 1–5.

[DR14] C. Dwork and A. Roth, “The algorithmic founda¬tions of differential privacy,” Found. Trends Theor. Com-put. Sci., vol. 9, no. 3–4, pp. 211–407, 2014. doi: 10.1561/0400000042.

[DR18] Y. Ding and B. D. Rao, “Dictionary learning-based sparse channel representation and estimation for FDD massive mimo systems,” IEEE Transactions on Wireless Communications, vol. 17, no. 8, pp. 5437–5451, 2018.

[Dua21] Q. Duan, "Intelligent and Autonomous Management in Cloud-Native Future Networks—A Survey on Related Standards from an Architectural Perspective." Future Internet 13, no. 2 (2021): 42.

[DZB+19] C. D’Andrea, A. Zappone, S. Buzzi, and M. Debbah, “Uplink power control in cell-free massive MIMO via deep learning,” in 2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2019, pp. 554–558.

[DZF+20] S. Deng, H. Zhao, W. Fang, J. Yin, S. Dustdar and A. Y. Zomaya, "Edge Intelligence: The Confluence of Edge Computing and Artificial Intelligence," in IEEE Internet of Things Journal, vol. 7, no. 8, pp. 7457-7469, Aug. 2020, doi: 10.1109/JIOT.2020.2984887.

[DZL+19] P. Dong, H. Zhang, G.Y. Li, I. Gaspar, N. Naderi-Alizadeh, “Deep CNN-Based Channel Estimation for mmWaveMassive MIMO Systems,”IEEE J. Select. Topics Signal Process.,Vol. 13, No. 5, pp. 989–1000, Sep. 2019.

[EAA20] Elsayed, M., Abdelkader, H., & Abdelwahab, A. (2020). Deep Learning Models for Heterogeneous Big Data Analytics. 2020 15th International Conference on Computer Engineering and Systems (ICCES), 1–5. https://doi.org/10.1109/ICCES51560.2020.9334569 , last accessed 16.06.2021

[EC19] “Ethic guidelines for trustworthy AI” https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[ED21] Exploit Database, https://www.exploit-db.com/

[EGE19] ETSI GS ENI: " Experiential Networked Intelligence (ENI); System Architecture ". Available: https://www.etsi.org/deliver/etsi_gs/ENI/001_099/005/01.01.01_60/gs_ENI005v010101p.pdf , last accessed 23.04.2021

[EGM18] ETSI GS MEC 002: "Multi-access Edge Computing (MEC); Phase 2: Use Cases and Requirements", v2.1.1, Oct. 2018.

[ENI21] Dahmen-Lhuissier, S. (n.d.). ETSI - Experiential Networked Intelligence (ENI). ETSI. Retrieved 23 April 2021, from https://www.etsi.org/technologies/experiential-networked-intelligence , last accessed 23.04.2021

[ERI19] “Edge computing and 5G” https://attom.tech/wp-content/uploads/2019/10/edge-computing-and-5g.pdf, June 2019

[Eri20] “Trustworthy AI: explainability, safety and verifiability” https://www.ericsson.com/en/blog/2020/12/trustworthy-ai

[FAL17] Chelsea Finn, Pieter Abbeel, Sergey Levine, “Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks”, Proceedings of the 34 th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

[FCD+18] A. Felix, S. Cammerer, S. Dörner, J. Hoydis, and S. Ten Brink, "OFDM-autoencoder for end-to-end learning of communications systems," in Proc. IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 2018.

[FDA17] Florensa, Carlos, Yan Duan, and Pieter Abbeel. "Stochastic neural networks for hierarchical reinforcement learning." arXiv preprint arXiv:1704.03012 (2017).

[FGC12] Frost, S. A., Goebel, K., & Celaya, J. (n.d.). A Briefing on Metrics and Risks for Autonomous Decision- Making in Aerospace Applications. Available: https://ntrs.nasa.gov/api/citations/20120014264/downloads/20120014264.pdf , last accessed 23.04.2021

[Flo02] L. Floridi, "what is the philosophy of information", Metaphilosophy, Vol. 33, pp 123-145, 2002.

[FPE17] X. Foukas, G. Patounas, A. Elmokashfi and M. K. Marina, "Network Slicing in 5G: Survey and Challenges," in IEEE Communications Magazine, vol. 55, no. 5, pp. 94-100, May 2017.

[FS20] H. Farhadi, M. Sundberg, "Context-aware receiver for high-band transmission," IEEE Global Communication Conference, Taipei, Taiwan, Dec. 2020.

[FSW19] R. Fritschek, R. F. Schaefer, et G. Wunder, « Deep Learning for Channel Coding via Neural Mutual Information Estimation », arXiv:1903.02865, 2019.

[FYB18] C. Fung, C. J. Yoon, and I. Beschastnikh, "Mitigating sybils in federated learning poisoning," CoRR, arXiv:1808.04866, 2018.

[GBC16] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, pp. 555–584. MIT Press, 2016. http://www.deeplearningbook.org .

[GBL+03] Ghahramani, Z. (2004). Unsupervised Learning. In O. Bousquet, U. von Luxburg, & G. Rätsch (Eds.), Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2—14, 2003, Tübingen, Germany, August 4—16, 2003, Revised Lectures (pp. 72–112). Springer. https://doi.org/10.1007/978-3-540-28650-9_5 , last accessed 16.06.2021

[GCH+17] T. Gruber, S. Cammerer, J. Hoydis, and S. t. Brink, "On deep learning based channel decoding," in Proceedings of the 51st Annual Conference on Information Sciences and Systems, pp. 1–6, 2017.

[GDPR] General Data Protection Regulation. Online: https://gdpr-info.eu/

[GGD+19] D.M. Gutierrez-Estevez, M. Gramaglia, A. De Domenico, G. Dandachi, S. Khatibi, D. Tsolkas, I. Balan, A. Garcia-Saavedra, U. Elzur, and Y. Wang, "Artificial intelligence for elastic management and orchestration of 5G networks," IEEE Wireless Communications, 26(5), pp.134-141, 2019.

[GH20] C.M. de Galland, C. Monnoyer, and J.M. Hendrickx. "Fundamental Performance Limitations for Average Consensus in Open Multi-AgentSystems." arXiv preprint arXiv:2004.06533 (2020).

[GJW+18] X. Gao, S. Jin, C. Wen, and G. Y. Li, "ComNet: Combination of deep learning and expert knowledge in OFDM receivers," IEEE Communications Letters, vol. 22, no. 12, pp. 2627–2630, Dec 2018.

[GLD+19] T. Gu, K. Liu, B. Dolan-Gavitt and S. Garg, "BadNets: Evaluating Backdooring Attacks on Deep Neural Networks," IEEE Access, vol. 7, pp. 47230-47244, 2019.

[GLSL16] Gu, S., Lillicrap, T., Sutskever, I., & Levine, S. (2016, June). Continuous deep Q-learning with model-based acceleration. In International Conference on Machine Learning (pp. 2829-2838). PMLR.

[Goo16] I. Goodfellow, Ian. "NIPS 2016 tutorial: Generative adversarial networks." arXiv preprint arXiv:1701.00160 (2016).

[GOP19] M. Gürbüzbalaban, A. Ozdaglar, and P.A. Parrilo. "Why random reshuffling beats stochastic gradient descent." Mathematical Programming (2019): 1-36.

[Guo20] W. Guo, "Explainable artificial intelligence for 6G: Improving trust between human and machine." IEEE Communications Magazine 58.6 (2020): 39-45.

[GYS18] B. Güler, A. Yener and A. Swami, "The semantic communication game", IEEE Trans. Cogn. Commun. Netw., 4 (2018), pp. 787-802.

[GYY+20] Geiger, R. Stuart, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. "Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?." In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 325-336. 2020.

[HAA20] He, C., Annavaram, M., & Avestimehr, S. (2020). Group Knowledge Transfer: Federated Learning of Large CNNs at the Edge. arXiv: Learning.

[HBA+20] S. Hosseinalipour, C. G. Brinton, V. Aggarwal, H. Dai and M. Chiang, "From Federated to Fog Learning: Distributed Machine Learning over Heterogeneous Wireless Networks," IEEE Communications Magazine, 58(12), pp. 41-47, Dec. 2020. 10.1109/MCOM.001.2000410.

[HBK+16] Hegedűs, I., Berta, Á., Kocsis, L., Benczúr, A. A., & Jelasity, M. (2016). Robust decentralized low-rank matrix decomposition. ACM Transactions on Intelligent Systems and Technology (TIST), 7(4), 1-24.

[HDW+19] C. Hellings, A. Dehmani, S. Wesemann, M. Koller, W, Utschick, "Evaluation of Neural-Network-Based Channel Estimators Using Measurement Data," in Proc. ITG Workshop on Smart Antennas (WSA'19), Apr. 2019.

[HFZ+19] L. Huang, X. Feng, C. Zhang, L. Qian and Y. Wu, "Deep reinforcement learning-based joint task offloading and bandwidth allocation for multi-user mobile edge computing," in Digital Communications and Networks 5, no. 1, pp. 10-17, 2019.

[HHH+20] F. Hussain, S. A. Hassan, R. Hussain and E. Hossain, "Machine learning for resource management in cellular and IoT networks: Potentials, current solutions, and open challenges," in IEEE Communications Surveys & Tutorials, vol. 22, no. 2, pp. 1251-1275, Second quarter 2020

[HJW+19] H. He, S. Jin, C. Wen, F. Gao, G. Y. Li, and Z. Xu, "Model-driven deep learning for physical layer communications," IEEE Wireless Communications, vol. 26, no. 5, pp. 77–83, 2019.

[HKH21] M. Honkala, D. Korpi, and J. M. J. Huttunen, "DeepRx: Fully convolutional deep learning receiver," IEEE Transactions on Wireless Communications, 2021.

[HKT19] P. Hernandez-Leal, B. Kartal, M. E. Taylor. A Survey and Critique of Multiagent Deep Reinforcement Learning. Autonomous Agents and Multi-Agent Systems, vol. 33, no. 6, pp. 1-48, 2019.

[HRW14] J. R. Hershey, J. L. Roux, and F. Weninger, “Deep unfolding: Model-based inspiration of novel deep architectures,” arXiv preprint, arXiv:1409.2574, 2014.

[HS97] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” IEEE Neural Comput., vol. 9, no. 8, p. 1735–1780, 1997.

[HWJ+18] H. He, C. Wen, S. Jin, and G. Li, “Deep learning-based channel estimation for beamspace mmWave massive MIMO systems,” IEEE Wireless Communications Letters, vol. 7, no. 5, 2018.

[HZL+19] L. Huang, H. Zhang, R. Li, Y. Ge, et J. Wang, « AI Coding: Learning to Construct Error Correction Codes », arXiv:1901.05719, 2019.

[HZY+20] L. Huang, L. Zhang, S. Yang, L. P. Qian and Y. Wu, "Meta-Learning based Dynamic Computation Task Offloading for Mobile Edge Computing Networks," in IEEE Communications Letters, Early Access, 2020.

[IAR21] I. Ismath, S. Ali, N. Rajatheva, and M. Latva-aho, “Deep Contextual Bandits for Fast Neighbor Aided Initial Access in mmWave Cell Free Networks,” 2021. [Online]. Available: https://arxiv.org/abs/2009.06974

[IBN+19] G. Interdonato, E. Björnson, H.Q. Ngo, P. Frenger, and E.G. Larsson, “Ubiquitous cell-free Massive MIMO communications”, EURASIP Journal on Wireless Communications and Networking volume 2019, Article number: 197 (2019).

[IMA21] I. Ismath, K. B. S. Manosha, S. Ali, N. Rajatheva, and M. Latva-aho, “Deep Contextual Bandits forFast Initial Access in mmWave Based User Centric Ultra Dense Networks.”IEEE Vehicular Technology Conference (accepted), 2021.. Available: https://arxiv.org/abs/2009.06974.2

[ISM01] M. Islam, Shahjahan, and K. Murase, “A new weight freezing method for reducing training time in designing artificial neural networks,” in 2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat.No.01CH37236), 2001, vol. 1, pp. 341–346 vol.1, doi: 10.1109/ICSMC.2001.969835.

[ITU20] ITU Focus Group on Machine Learning for Future Networks including 5G. Online: https://www.itu.int/en/ITU-T/focusgroups/ml5g/Pages/default.aspx

[JC20] T.L. Jensen and E. De Carvalho, “An optimal channel estimation scheme for intelligent reflecting surfaces based on a minimum variance unbiased estimator,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE,2020, pp. 5000–5004.

[JRL19] N. Jayaweera, N. Rajatheva, and M. Latvaaho, “Autonomous driving without a burden: View from outside with elevated lidar,” in 2019 IEEE 89th Vehicular Technology Conference (VTC 2019 Spring), 2019, pp. 1–7.

[KBG+18] A. Klautau, P. Batista, N. González-Prelcic, Y. Wang and R. W. Heath, "5G MIMO Data for Machine Learning: Application to Beam-Selection Using Deep Learning," 2018 Information Theory and Applications Workshop (ITA), 2018, pp. 1-9.

[KGH19] A. Klautau, N. González Prelcic, and R. W. Heath, “Lidar data for deep learning based mmwave beamselection,” IEEE Wireless Communications Letters, vol. 8, no. 3, pp. 909–912, 2019.

[KHH+21] D. Korpi, M. Honkala, J. M. J. Huttunen, and V. Starck, “DeepRx MIMO: Convolutional MIMO detection with learned multiplicative transformations,” in Proc. IEEE International Conference on Communications (ICC), 2021.

[KHU19] M. Koller, C. Hellings, W. Utschick, “Learning-Based ChannelEstimation for Various Antenna Array Configurations,” in Proc. IEEE20th Int. Workshop Signal Process. Advances Wireless Commun.(SPAWC’19), Jul. 2019.

[KJR+18] H. Kim, Y. Jiang, R. Rana, S. Kannan, S. Oh, and P. Viswanath, “Communication Algorithms via Deep Learning,” ArXiv180509317 Cs Stat, 2018.

[KM+21] P. Kairouz and B.H. McMahan (2021), "Advances and Open Problems in Federated Learning", Foundations and Trends® in Machine Learning: Vol. 14: No. 1. http://dx.doi.org/10.1561/2200000083

[KM21] N. K. Kundu and M. R. McKay, "Channel estimation for reconfigurable intelligent surface aided MISO communications: From LMMSE to deep learning solutions," in IEEE Open Journal of the Communications Society, vol. 2, pp. 471-487, 2021, doi: 10.1109/OJCOMS.2021.3063171.

[KMH+19] D. Kim, S. Moon, D. Hostallero, W.J. Kang, T. Lee, K. Son, and Y. Yi, "Learning to schedule communication in multi-agent reinforcement learning," in International Conference on Learning Representations (ICLR), 2019.

[KMR+16] J. Konečný, B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence." arXiv preprint arXiv:1610.02527 (2016).

[Koh12] J. Kohlas, “Information Algebras: Generic Structures for Inference”, Springer Science & Business Media. 2012.

[KOT07] S. B. Kotsiantis. Supervised Machine Learning | Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies. (n.d.)., from https://dl.acm.org/doi/10.5555/1566770.1566773 , last accessed 16.06.2021

[KPR17] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell, “Overcoming catastrophic forgetting in neural networks”

[Kra91] M. A. Kramer, “Nonlinear principal component analysis using autoassociative neural networks,” AIChE Journal, vol. 37, no. 2, pp. 233–243, 1991, doi:https://doi.org/10.1002/aic.690370209.

[KW13] D.P. Kingma, and M. Welling, “Auto-encoding variational bayes”. arXiv preprint arXiv:1312.6114, 2013.

[LBH15] Y. LeCun, Y. Bengio, and G. E. Hinton, “Deep learning,” Nature, vol.521, no. 7553, pp. 436–444, 2015.

[LCH+19] M. Lian, F. Carpi, C. Häger , and H. D. Pfister, “Learned belief propagation decoding with simple scaling and snr adaptation,” in Proceedings of the 2019 IEEE International Symposium on Information Theory, pp. 161–165, 2019.

[LDD+20] G. Larue, , M. Dhiflaoui, L.A. Dufrene, Q. Lampin, P. Chollet, H. Ghauch, and G. Rekaya, "Low-Complexity Neural Networks for Baseband Signal Processing," in 2020 IEEE Globecom Workshops (GC Wkshps, pp. 1-6. IEEE, 2020.doi:10.1109/GCWkshps50303.2020.9367521.

[LDL+21] G. Larue, L.A. Dufrene, Q. Lampin, P. Chollet, H. Ghauch and G. Rekaya, "Blind Neural Belief Propagation Decoder for Linear Block Codes", to appear in the proceedings of 2021 EuCNC & 6G Summit – PHY, 2021.

[LEM20] Max Lemke, Next generation of IoT, Opportunities for Europe, https://aioti.eu/wp-content/uploads/2020/09/200929-Next-Generation-IoT-M-Lemke.pdf, September 2020.

[LeM21a] L. Le Magoarou, "Similarity-based prediction for channel mapping and user positioning," in IEEE Communications Letters, 2021.

[LeM21b] L. Le Magoarou, "Efficient channel charting via phase-insensitive distance computation," (work in progress), arXiv:2104.13184, 2021.

[LGJ1-20] H. Lee, M. Girnyk, and J. Jeong, "Deep MIMO Autoprecoder," in Proceedings of IEEE International Conference on Communications, June 2020.

[LGJ2-20] H. Lee, M. Girnyk, and J. Jeong, "Deep reinforcement learning approach to MIMO precoding problem: Optimality and Robustness," https://arxiv.org/abs/2006.16646, June 2020.

[LJC+00] J. Li, J., J. Jannotti, J., D.S. De Couto, D. S., D.R. Karger, D. R., and& R. Morris, R. (2000, August). A scalable location service for geographic ad hoc routing. In/Proceedings of the 6th annual international conference on Mobile computing and networking/ (pp. 120-130). ACM.

[LLH+20] W.Y.B. Lim, N.C. Luong, D.T. Hoang, Y. Jiao, Y.C. Liang, Q. Yang, D. Niyato, and C. Miao. "Federated learning in mobile edge networks: A comprehensive survey." IEEE Communications Surveys & Tutorials 22, no. 3 (2020): 2031-2063.

[LLV20] Li, Chengxi & Li, Gang & Varshney, Pramod. (2020). Decentralized Federated Learning via Mutual Knowledge Transfer.

[LMZ+20] H. Liao, Y. Mu, Z. Zhou, M. Sun, Z. Wang and C. Pan, "Blockchain and Learning-Based Secure and Intelligent Task Offloading for Vehicular Fog Computing," in IEEE Transactions on Intelligent Transportation Systems, 2020.

[LSY+20] F. Liang, C. Shen, W. Yu, and F. Wu, "Towards optimal power control via ensembling deep neural networks," IEEE Transactions on Communications, vol. 68, no. 3, pp. 1760–1776, 2020.

[LSZ+19] Liu, Z, Sun, M, Zhou, T, et al. Rethinking the value of network pruning. In: International conference on learning representations (ICLR), New Orleans, 5–6 May 2019, pp.1–21.

[LV19] Y. Li and N. Vasconcelos, "REPAIR: Removing Representation Bias by Dataset Resampling," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9564-9573.

[LWP17] R. Lowe, Y. Wu, Pritzel, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in

Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS), pp. 6382–6393, December 2017.

[LXK+20] W. Y. B. Lim et al., "When Information Freshness Meets Service Latency in Federated Learning: A Task-Aware Incentive Scheme for Smart Industries," in IEEE Transactions on Industrial Informatics, 2020.

[LZJ+20] H. Liu, J. Zhang, S. Jin and B. Ai, "Graph Coloring Based Pilot Assignment for Cell-Free Massive MIMO Systems," in IEEE Transactions on Vehicular Technology, vol. 69, no. 8, pp. 9180-9184, Aug. 2020

[LZZ+19] D. Liu, G. Zhu, J. Zhang, and K. Huang, July. “Wireless data acquisition for edge learning: Importance-aware retransmission”, In International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), pp. 1-5, 2019.

[MAR10] T.L. Marzetta, “Noncooperative Cellular Wireless with UnlimitedNumbers of Base Station Antennas,”IEEE Trans. Wireless Commun.,Vol. 9, No. 11, pp. 3590–3600, Nov. 2010.

[MBP+10] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “Online learning for matrix factorization and sparse coding,” Journal of Machine Learning Research, vol. 11, no. 1, pp. 19–60, Jan. 2010.

[MCC+19] F. Mattioli, D. Caetano, A. Cardoso, E. Naves, and E. Lamounier, "An experiment on the use of genetic algorithms for topology selection in deep learning," Journal of Electrical and Computer Engineering, 2019.

[MEA20] F. B. Mismar, B. L. Evans and A. Alkhateeb, "Deep reinforcement learning for 5G networks: Joint beamforming, power control, and interference coordination," in IEEE Transactions on Communications, vol. 68, no. 3, pp. 1581-1592, March 2020, doi: 10.1109/TCOMM.2019.2961332.

[MER21] M. Merluzzi, P. D. Lorenzo and S. Barbarossa, "Wireless Edge Machine Learning: Resource Allocation and Trade-Offs," in IEEE Access, vol. 9, pp. 45377-45398, 2021, doi: 10.1109/ACCESS.2021.3066559.

[MH19] Monnoyer de Galland, Charles, and Julien M. Hendrickx. "Lower bound performances for average consensus in open multi-agent systems (extended version)." arXiv e-prints (2019): arXiv-1909.

[Mit21] MITRE Common Vulnerabilities & Exposures, https://cve.mitre.org/

[MJ19] D. Mishra and H. Johansson, “Channel estimation and low-complexity beamforming design for passive intelligent surface assisted MISO wireless energy transfer,” inICASSP 2019 - 2019 IEEE International Conferenceon Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 4659–4663.

[MMR+17] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Artificial Intelligence and Statistics, pp. 1273-1282. PMLR, 2017.

[MP20] L. Le Magoarou and S. Paquelet, “Online unsupervised deep unfolding for massive mimo channel estimation,” arXiv preprint arXiv:2004.14615, 2020.

[MSD+19] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, “Exploiting unintended feature leakage in collaborative learning,” in SP, 2019, pp. 691–706.

[Nad21] N. Naderializadeh, "Contrastive Self-Supervised Learning for Wireless Power Control," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.

[Nag19] Anudit Nagar (2019) Privacy-Preserving Blockchain Based Federated Learning with Differential Data Sharing, arXiv:1912.04859 [cs.CR]

[NAY17] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell Free Massive MIMO Versus Small Cells," IEEE Transactions on Wireless Communications, vol. 16, no. 3, pp. 1834–1850, 2017.

[NC02] M.A. Nielsen and I. Chuang, "Quantum computation and quantum information", 2002.

[NCF18] NIST Cybersecurity Framework, https://www.nist.gov/cyberframework

[NDR20] Niknam, S., Dhillon, H. S., & Reed, J. H. (2020). Federated Learning for Wireless Communications: Motivation, Opportunities, and Challenges. IEEE Communications Magazine, 58(6), 46–51. https://doi.org/10.1109/MCOM.001.1900461 , last accessed 16.06.2021

[NG19] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," in IEEE Journal on Selected Areas in Communications, vol. 37, no. 10, pp. 2239-2250, Oct. 2019, doi: 10.1109/JSAC.2019.2933973.

[NML+18] E. Nachmani, E. Marciano, L. Lugosch, W. J. Gross, D. Burshtein, et Y. Be'ery, "Deep Learning Methods for Improved Decoding of Linear Codes", IEEE Journal of selected topic in signal processing, 2018.

[NML18] E. Nachmani, E. Marciano, L. Lugosch, W. J. Gross, D. Burshtein, and Y. Be'ery, "Deep learning methods for improved decoding of linear codes," IEEE Journal of Selected Topics in Signal Processing, vol. 12, no. 1, pp. 119–131, 2018.

[NMRG] IETF Network Management Research Group, "Digital Twin Network: Concepts and Reference Architecture". Online: https://datatracker.ietf.org/doc/draft-zhou-nmrg-digitaltwin-network-concepts/

[NPC] NIST Privacy-Enhancing Cryptography https://csrc.nist.gov/projects/pec

[NPE] NIST Privacy Engineering Program, https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering

[NSM] https://docs.rocknsm.io/reference/tutorials/

[NWU18] D. Neumann, T. Wiese, and W. Utschick, "Learning the MMSE channel estimator," IEEE Transactions on Signal Processing, vol. 66, no. 11, pp. 2905–2917, June 2018.

[OAIDL18] Artificial Intelligence and Deep Learning—Developer Wiki—Confluence. (n.d.)., from https://wiki.onap.org/display/DW/Artificial+Intelligence+and+Deep+Learning , last accessed 16.06.2021

[OC16] J. Oueis and E. Calvanese Strinati, "Uplink Traffic in Future Mobile Networks: Pulling the Alarm", CrownCom 2016: Cognitive Radio Oriented Wireless Networks pp 583-593.

[OH17] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," in IEEE Transactions on Cognitive Communications and Networking, vol. 3, no. 4, pp. 563-575, Dec. 2017, doi: 10.1109/TCCN.2017.2758370.

[ONA21] ONAP web site, available: https://docs.onap.org/en/latest/ , last accessed 08.04.2021.

[ORW+18] T. J. O'Shea, T. Roy, N. West, and B. C. Hilburn, "Physical layer communications system design over-the-air using adversarial networks," in 2018 26th Eur. Signal Process. Conf. (EUSIPCO), Sep. 2018, pp. 529–532.

[ORW19] T. J. O'Shea, T. Roy and N. West, "Approximating the Void: Learning Stochastic Channel Models from Observation with Variational Generative Adversarial Networks," 2019 International Conference on Computing, Networking and Communications (ICNC), Honolulu, HI, USA, 2019, pp. 681-686, doi: 10.1109/ICCNC.2019.8685573.

[OSM20] OSM. (n.d.). https://osm.etsi.org , last accessed 06.06.2021

[OSM21] OSM PoC 11 Deployment of AI-Agents in OSM - OSM Public Wiki. (n.d.). Available: https://osm.etsi.org/wikipub/index.php/OSM_PoC_11_Deployment_of_AI-Agents_in_OSM , last accessed 23.04.2021

[OSMRT20] OSM_Release_TEN_-_Release_Notes.pdf. (n.d.). Available: https://osm.etsi.org/wikipub/images/3/30/OSM_Release_TEN_-_Release_Notes.pdf , last accessed 23.04.2021

[PBC+20] E. Peltonen, M. Bennis, M. Capobianco, M. Debbah, A. Ding, F. Gil-Castiñeira, M. Jurmu, T. Karvonen, M. Kelanti, A. Kliks, and T. Leppänen, "6G white paper on edge intelligence". arXiv preprint arXiv:2004.14850. 2020.

[PBC+20] Peltonen, E., Bennis, M., Capobianco, M., Debbah, M., Ding, A., Gil-Castiñeira, F., Jurmu, M., Karvonen, T., Kelanti, M., Kliks, A., Leppänen, T., Lovén, L., Mikkonen, T., Rao, A., Samarakoon, S., Seppänen, K., Sroka, P., Tarkoma, S., & Yang, T. (2020). 6G White Paper on Edge Intelligence [White paper]. (6G Research Visions, No. 8). University of Oulu. http://urn.fi/urn:isbn:9789526226774 , last accessed 16.06.2021.

[PEE21] Five Disruptive Features of Tomorrow's 6G Networks | Network Computing. (n.d.)., from https://www.networkcomputing.com/wireless-infrastructure/five-disruptive-features-tomorrow%E2%80%99s-6g-networks , last accessed 16.06.2021.

[PMI21] M. Padmal, D. Marasinghe, V. Isuru, N. Jayaweera, S. Ali, and N. Rajatheva, "Elevated lidar based sensing for 6g – 3d maps with cm level accuracy," 2021. Available: https://arxiv.org/abs/2102.10849

[PML+19] J. Portilla, G. Mujica, J. -S. Lee and T. Riesgo, "The Extreme Edge at the Bottom of the Internet of Things: A Review," in IEEE Sensors Journal, vol. 19, no. 9, pp. 3179-3190, 1 May1, 2019, doi: 10.1109/JSEN.2019.2891911.

[PPK+21] S. R. Pokhrel, L. Pan, N. Kumar, R. Doss and H. Le Vu, "Multipath TCP Meets Transfer Learning: A Novel Edge-based Learning For Industrial IoT," in IEEE Internet of Things Journal, Early Access, 2021.

[PSB+19] J. Park, S. Samarakoon, M. Bennis and M. Debbah, "Wireless Network Intelligence at the Edge," in Proceedings of the IEEE, vol. 107, no. 11, pp. 2204-2239, Nov. 2019, doi: 10.1109/JPROC.2019.2941458.

[RAD78] R. L. Rivest, L. Adleman, and M. L. Dertouzos. “On data banks and privacy homomorphisms.” In Foundations of Secure Computation, 1978.

[RB20] B. Rassouli and D. Gündüz, "Optimal Utility-Privacy Trade-Off With Total Variation Distance as a Privacy Measure," in IEEE Transactions on Information Forensics and Security, vol. 15, pp. 594-603, 2020.

[RED] Radio Equipment Directive. Online: https://ec.europa.eu/growth/sectors/electrical-engineering/red-directive_en

[RHW+20] J. Ren, Y. He, D. Wen, G. Yu, K. Huang and D. Guo, "Scheduling for Cellular Federated Edge Learning With Importance and Channel Awareness," in IEEE Transactions on Wireless Communications, vol. 19, no. 11, pp. 7690-7703, Nov. 2020.

[RMJ+20] A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, et J. Kepner, "Survey of Machine Learning Accelerators", arXiv:2009.00993v1, 2020.

[RMR+20] N. Rajapaksha, K. B. S. Manosha, N. Rajatheva, and M. Latva-aho, “Deep learning-based power control for cell-free massive MIMO networks”, 2020. [Online]. Available: https://arxiv.org/pdf/2102.10366.pdf

[RPB+20] A. Rago, G. Piro, G. Boggia and P. Dini, "Multi-Task Learning at the Mobile Edge: An Effective Way to Combine Traffic Classification and Prediction," in IEEE Transactions on Vehicular Technology, vol. 69, no. 9, pp. 10362-10374, Sept. 2020.

[RS14] N. Raveendran et S. G. Srinivasa, "An analysis into the loopy belief propagation algorithm over short cycles ", in 2014 IEEE International Conference on Communications (ICC), Sydney, NSW, June 2014, p. 2009-2014

[RTM+19] A. Reisizadeh, A., H. Taheri, H., A. Mokhtari,, A., H. Hassani, H., and R.& Pedarsani, R. Advances in Neural Information Processing Systems 32 (NIPS 2019) (2019).

[RWT+18] Riazi, M., Weinert, C., Tkachenko, O., Songhori, E.M., Schneider, T., & Koushanfar, F. Chameleon: A Hybrid Secure Computation Framework for Machine Learning Applications. Proceedings of the 2018 on Asia Conference on Computer and Communications Security.

[S15] Strom, Nikko. "Scalable distributed DNN training using commodity GPU cloud computing." Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[Sal+19] V. Salapura et al., "Generative Policy Framework for AI Training Data Curation," 2019 IEEE International Conference on Smart Computing (SMARTCOMP), 2019, pp. 475-477.

[SB17] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction. Second Edition, MIT Press, Cambridge, Massachusetts, London, 2017.

[SBL+18] R. Sharma, S. Biookaghazadeh, B. Li, and M. Zhao, “Are Existing Knowledge Transfer Techniques Effective for Deep Learning with Edge Devices?,” in 2018

IEEE International Conference on Edge Computing (EDGE), 2018, pp. 42–49, doi:10.1109/EDGE.2018.00013.

[SDW19] N. Samuel, T. Diskin, and A. Wiesel, “Learning to detect,” IEEE Transactions on Signal Processing, vol. 67, no. 10, pp. 2554–2564, May 2019.

[SH19] O. Shental and J. Hoydis, “”Machine LLRning”: Learning to softly demodulate,” in Proc. IEEE Globecom Workshops (GC Wkshps), 2019.

[SKA19] N. Skatchkovsky and O. Simeone, ‘‘Optimizing pipelined computation and communication for latency-constrained edge learning,’’IEEE Commun.Lett., vol. 23, no. 9, pp. 1542–1546, Sep. 2019.

[SLC+20] R. Shafin, L. Liu, V. Chandrasekhar, H. Chen, J. Reed and J. C. Zhang, "Artificial Intelligence-Enabled Cellular Networks: A Critical Path to Beyond-5G and 6G," in IEEE Wireless Communications, vol. 27, no. 2, pp. 212-217, April 2020, doi: 10.1109/MWC.001.1900323.

[SMG+18] C. Studer, S. Medjkouh, E. Gönültas¸, T. Goldstein, and O. Tirkkonen, “Channel charting: Locating users within the radio environment using channel state information,” IEEE Access, vol. 6, pp. 47 682–47 698, 2018.

[SPW18] R. Sabbagh, C. Pan and J. Wang, "Pilot Allocation and Sum-Rate Analysis in Cell-Free Massive MIMO Systems," 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, USA, 2018, pp. 1-6

[SSS+17] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, “Membershipinference attacks against machine learning models,” in SP, 2017, pp. 3–18.

[STG+20] S. Sevgican, M. Turan, K. Gökarslan, H. B. Yilmaz and T. Tugcu, "Intelligent network data analytics function in 5G cellular networks using machine learning," in Journal of Communications and Networks, vol. 22, no. 3, pp. 269-280, Jun. 2020.

[SW95]

[SWM17] W. Samek, T. Wiegand, and K.L. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models." arXiv preprint arXiv:1708.08296 (2017).

[SZD18] L. Sanguinetti, A. Zappone, and M. Debbah, “Deep learning power allocation in massive MIMO,” in 2018 52nd Asilomar Conference on Signals, Systems, and Computers, 2018, pp. 1257–1261.

[SZHT07] Swami, A., Zhao, Q., Hong, Y. W., & Tong, L. (Eds.). (2007). Wireless sensor networks: signal processing and communications perspectives. John Wiley & Sons.

[TFF21] TensorFlow Federated, https://github.com/tensorflow/federated

[TGK+19] A. Tavanaei, M. Ghodrati, S.R. Kheradpisheh, T. Masquelier, A. Maida, "Deep learning in spiking neural networks," in Neural Netw. 2019 Mar;111:47-63. doi: 10.1016/j.neunet.2018.12.002.

[TH09] Issariyakul, Teerawat, and Ekram Hossain. "Introduction to network simulator 2 (NS2)." Introduction to network simulator NS2. Springer, Boston, MA, 2009. 1-18.

[TL20] Y. Tao, and S. Lu. "From Online to Non-iid Batch Learning." Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020.

[TSW+21] N. Truong, K. Sun, S. Wang, F. Guitton, Y. Guo, “Privacy preservation in federated learning: An insightful survey from the GDPR perspective”, Computers & Security, Volume 110, 2021, 102402, ISSN 0167-4048, https://doi.org/10.1016/j.cose.2021.102402.

[TZJ+16] F. Tramer, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, “Stealing machine learning models via prediction apis,” CoRR, vol. abs/1609.02943, 2016.

[VNB+20] T. Van Chien, T. Nguyen Canh, E. Björnson, and E. G. Larsson, “Power control in cellular massive MIMO with varying user activity: A deep learning solution,” IEEE Transactions on Wireless Communications, vol. 19, no. 9, pp. 5732–5748, 2020.

[VWK+20] J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen, and J.S. Rellermeyer. 2020, "A Survey on Distributed Machine Learning," in ACM Comput. Surv. 53, 2, Article 30, Jun. 2020.

[WD92] C. J. C. H. Watkins and P. Dayan, “Technical note: Q-learning,” Machine Learning, vol. 8, pp. 279–292, 1992.

[WFC18] Wang, Y., Forbes, R., Cavigioli, C., Wang, H., Gamelas, A., Wade, A., Strassner, J., Cai, S., & Liu, S. (2018). Network Management and Orchestration Using Artificial Intelligence: Overview of ETSI ENI. IEEE Communications Standards Magazine, 2(4), 58–65. https://doi.org/10.1109/MCOMSTD.2018.1800033 , last accessed 23.04.2021

[WFH+18] Wei, L., Foh, C. H., He, B., & Cai, J. (2018). Towards Efficient Resource Allocation for Heterogeneous Workloads in IaaS Clouds. IEEE Transactions on Cloud Computing, 6(1), 264–275. https://doi.org/10.1109/TCC.2015.2481400 , last accessed 16.06.2021

[WG20] A. Willner and V. Gowtham, "Toward a Reference Architecture Model for Industrial Edge Computing," in IEEE Communications Standards Magazine, vol. 4, no. 4, pp. 42-48, December 2020, doi: 10.1109/MCOMSTD.001.2000007.

[WiKi+21] Wikipedia contributors. (2021, July 12). Autoregressive–moving-average model. In Wikipedia, The Free Encyclopedia. Retrieved 17:33, July 16, 2021, from https://en.wikipedia.org/w/index.php?title=Autoregressive%E2%80%93moving-average_model&oldid=1033303419

[WK05] F.M. Willems and T. Kalker, “Semantic compaction, transmission, and compression codes”, Proceedings. International Symposium on Information Theory, ISIT 2005, IEEE (2005), pp. 214-218.

[WLT+21] M. Wang, Y. Lin, Q. Tian and G. Si, "Transfer Learning Promotes 6G Wireless Communications: Recent Advances and Future Challenges," in IEEE Transactions on Reliability, Early Access, 2021.

[WLZ+19] D. Wen, X. Li, Q. Zeng, J. Ren and K. Huang, "An Overview of Data-Importance Aware Radio Resource Management for Edge Machine Learning," in Journal of Communications and Information Networks, vol. 4, no. 4, pp. 1-14, Dec. 2019.

[WPC+21] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang and P. S. Yu, "A Comprehensive Survey on Graph Neural Networks," in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 1, pp. 4-24, Jan. 2021.

[WSC21] D. Wilms, C. Stoecker and J. Caballero, "Data Provenance in Vehicle Data Chains," 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring), 2021, pp. 1-5.

[WSR+20] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-y. Sohn, K. Lee, and D. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning,"NeurIPS, 2020.

[WTK17] Wojciech Samek, Thomas Wiegand, Klaus-Robert Müller, " Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models". arXiv:1708.08296

[WZZ+ 21] Q. Wu, S. Zhang, B. Zheng, C. You and R. Zhang, "Intelligent Reflecting Surface Aided Wireless Communications: A Tutorial," in IEEE Transactions on Communications., Early Access, 2021

[WZZ+1811] Y. Wang, Z. Zhang, S. Zhang, S. Cao, and S. Xu, "A Unified Deep Learning Based Polar-LDPC Decoder for 5G Communication Systems," in 2018 10th International Conference on Wireless Communications and Signal Processing (WCSP), Hangzhou, 2018, pp. 1–6.

[XGJ20] W. Xu, F. Gao, S. Jin, and A. Alkhateeb, "3d scene based beam selection for mmwave communications,"2019. Available: https://arxiv.org/abs/1911.08409

[XMT17] Xiao, Kefan, Shiwen Mao, and Jitendra K. Tugnait. "MAQ: A multiple model predictive congestion control scheme for cognitive radio networks." IEEE Transactions on Wireless Communications 16.4 (2017): 2614-2626.

[XQL+20] H. Xie, Z. Qin, G.Y. Li, and B.H. Juang. "Deep learning enabled semantic communication systems." arXiv preprint arXiv:2006.10685 (2020).

[Yao82] Yao, A.. "Protocols for secure computations." FOCS 1982 (1982).

[YCX+20] Yang, B., Cao, X., Xiong, K., Yuen, C., Guan, Y. L., Leng, S., Qian, L., & Han, Z. (2020). Edge Intelligence for Autonomous Driving in 6G Wireless System: Design Challenges and Solutions. ArXiv:2012.06992 [Cs]. http://arxiv.org/abs/2012.06992 , last accessed 16.06.2021

[YLJ+18] H. Ye, G. Y. Li, B. F. Juang and K. Sivanesan, "Channel Agnostic End-to-End Learning Based Communication Systems with Conditional GAN," 2018 IEEE Globecom Workshops (GC Wkshps), Abu Dhabi, United Arab Emirates, 2018, pp. 1-5, doi: 10.1109/GLOCOMW.2018.8644250.

[YLJ18] H. Ye, G. Y. Li, and B.-H. Juang, "Power of deep learning for channel estimation and signal detection in OFDM systems," IEEE Communications Letters, vol. 7, no. 1, pp. 114–117, Feb. 2018.

[YLL+20] H. Ye, L. Liang, G. Y. Li and, B. H. F. Juang, "Deep learning based end-to-end wireless communication systems with conditional GAN as unknown channel", in IEEE Transactions on Wireless Communications, vol. 19, no. 5, pp. 3133-3143, May 2020, doi: 10.1109/TWC.2020.2970707.

[YM20] T. Yassine and L. Le Magoarou, "mpNet: variable depth unfolded neural network for massive MIMO channel estimation". arXiv preprint arXiv:2008.04088, 2020.

[YU19] C. Yinka-Banjo, and O. Ugot. "A review of generative adversarial networks and its application in cybersecurity." Artificial Intelligence Review (2019): 1-16.https://doi.org/10.1007/s10462-019-09717-4

[YWH+21] X. You, C.X. Wang, J. Huang, et al. "Towards 6G wireless communication networks: vision, enabling technologies, and new paradigm shifts", Sci. China Inf. Sci. 64, 110301 (2021).

[ZAA21] Y. Zhang, M. Alrabeiah, and A. Alkhateeb, “Reinforcement Learning for Beam Pattern Design in Millimeter Wave and Massive MIMO Systems,” 2021. [Online]. Available: http://arxiv.org/abs/2102.09084

[ZCL+19a] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, “Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing,” Proceedings of the IEEE, vol. 107, no. 8, pp. 1738–1762, 2019, doi:10.1109/JPROC.2019.2918951.

[ZCL+19b] Jiayi Zhang, Shuaifei Chen, Yan Lin, JiaKang Zheng, Bo Ai, Lajos Hanzo, IEEE Access, Volume 7, 2019.

[ZDC+21] Zheng, W., Deng, R., Chen, W., Popa, R.A., Panda, A., & Stoica, I. (2021). Cerebro: A Platform for Multi-Party Cryptographic Collaborative Learning.

[ZLD+20] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, “Toward an intelligent edge: wireless communication meets machine learning”, IEEE Communications Magazine, vol. 58, no. 1, pp.19-25, 2020.

[ZLT+21] Z. Zeng, Y. Liu, W. Tang and F. Chen, "Noise Is Useful: Exploiting Data Diversity for Edge Intelligence," in IEEE Wireless Communications Letters, vol. 10, no. 5, pp. 957-961, May 2021.

[ZND20] Y. Zhao, I. G. Niemegeers, and S. H. De Groot, “Power allocation in cell-free massive MIMO: A deep learning method,” IEEE Access, vol. 8, pp. 87 185–87 200, 2020.

[ZSM+18] Y. Zhao, Y., I. Shumailov, I., R. Mullins, R., and& R. Anderson, R. (2018). To compress or not to compress: Understanding the interactions between adversarial attacks and neural network compression. arXiv preprint arXiv:1810.00208.

[ZSM21] Dahmen-Lhuissier, S. (n.d.). ETSI - ZSM - Zero touch network & Service Management. ETSI. Retrieved. https://www.etsi.org/technologies/zero-touch-network-service-management , last accessed 16.06.2021

[ZVF+20] V. Ziegler, H. Viswanathan, H. Flinck, M. Hoffmann, V. Räisänen and K. Hätönen, "6G Architecture to Connect the Worlds," in IEEE Access, vol. 8, pp. 173508-173520, 2020, doi: 10.1109/ACCESS.2020.3025032., 2020.

[ZVG+18] Z. Zhao, M. C. Vuran, F. Guo, and S. Scott, “Deep-waveform: A learned OFDM receiver based on deep complex convolutional networks,” 2018. [Online]. Available: https://arxiv.org/abs/1810.07181

[ZWH20] G. Zhu, Y. Wang and K. Huang, "Broadband Analog Aggregation for Low-Latency Federated Edge Learning," in IEEE Transactions on Wireless Communications, vol. 19, no. 1, pp. 491-506, Jan. 2020,

[ZZ19] Beixiong Zheng and Rui Zhang, “Intelligent reflecting surface-enhanced OFDM: Channel estimation and reflection optimization, “IEEE Wireless Communications Letters, vol. 9, no. 4, pp. 518–522, 2019.

[ZZL+19] K. Zhang, Y. Zhu, S. Leng, Y. He, S. Maharjan and Y. Zhang, "Deep Learning Empowered Task Offloading for Mobile Edge Computing in Urban Informatics," in IEEE Internet of Things Journal, vol. 6, no. 5, pp. 7635-7647, Oct. 2019.