SCIENCE CHINA
Information Sciences

• REVIEW •

July 2025, Vol. 68, Iss. 7, 171301:1–171301:61
https://doi.org/10.1007/s11432-024-4337-1

Overview of AI and communication for 6G network: fundamentals, challenges, and future research opportunities

Qimei CUI[1], Xiaohu YOU[2*], Ni WEI[3], Guoshun NAN[1*], Xuefei ZHANG[1], Jianhua ZHANG[4], Xinchen LYU[1], Ming AI[5], Xiaofeng TAO[1], Zhiyong FENG[4], Ping ZHANG[4], Qingqing WU[6], Meixia TAO[7], Yongming HUANG[2], Chongwen HUANG[8], Guangyi LIU[9], Chenghui PENG[10], Zhiwen PAN[2], Tao SUN[9], Dusit NIYATO[11], Tao CHEN[12], Muhammad Khurram KHAN[13], Abbas JAMALIPOUR[14], Mohsen GUIZANI[15] & Chau YUEN[16]

[1]*National Engineering Research Center of Mobile Network Technologies, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[2]*National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China*
[3]*School of Information Science and Engineering, Fudan University, Shanghai 200433, China*
[4]*School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[5]*CICT Mobile Communication Technology Co., Ltd., Beijing 100020, China*
[6]*Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*
[7]*Department of Electronic Engineering and Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai 200240, China*
[8]*College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310058, China*
[9]*China Mobile Research Institute, Beijing 100053, China*
[10]*Huawei Technologies, Shanghai 201206, China*
[11]*College of Computing and Data Science, Nanyang Technological University, Singapore 639798, Singapore*
[12]*VTT Technical Research Centre of Finland Ltd., Espoo FI-02044, Finland*
[13]*Center of Excellence in Information Assurance, King Saud University, Riyadh 11362, Saudi Arabia*
[14]*School of Electrical and Information Engineering, University of Sydney, Sydney NSW 2006, Australia*
[15]*Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi 99163, UAE*
[16]*School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore 639798, Singapore*

Received 6 December 2024/Revised 5 February 2025/Accepted 10 March 2025/Published online 2 April 2025

Abstract With the growing demand for seamless connectivity and intelligent communication, the integration of artificial intelligence (AI) and sixth-generation (6G) communication networks has emerged as a transformative paradigm. By embedding AI capabilities across various network layers, this integration enables optimized resource allocation, improved efficiency, and enhanced system robust performance. This paper presents a comprehensive overview of AI and communication for 6G networks, with a focus on their foundational principles, inherent challenges, and future research opportunities. We first review the integration of AI and communications in the context of 6G, exploring the driving factors behind incorporating AI into wireless communications, as well as the vision for the convergence of AI and 6G. The discourse then transitions to a detailed exposition of the envisioned integration of AI within 6G networks, divided into three progressive stages. The first stage, AI for network, focuses on employing AI to augment network performance, optimize efficiency, and enhance user service experiences. The second stage, network for AI, highlights the role of the network in facilitating and buttressing AI operations and presents key enabling technologies. We compare wireless network large models with conventional large language models (LLMs), and identify key design principles and components for building wireless network architectures. In the final stage, AI as a service, it is anticipated that future 6G networks will innately provide AI functions as services, supporting application scenarios like immersive communication and intelligent industrial robots. Specifically, we define the quality of AI service, which refers to a framework for measuring AI services within the network. We further summarize the standardization process of AI for wireless networks, highlighting key milestones and ongoing efforts. In addition, we analyze the critical challenges faced by the integration of AI and communications in 6G. Finally, we outline promising future research opportunities that are expected to drive the development and refinement of AI and 6G communications.

Keywords 6G, AI, AI and communication, AI for network, AI as a service, LLMs, network for AI

Citation Cui Q M, You X H, Wei N, et al. Overview of AI and communication for 6G network: fundamentals, challenges, and future research opportunities. Sci China Inf Sci, 2025, 68(7): 171301, https://doi.org/10.1007/s11432-024-4337-1

* Corresponding author (email: xhyu@seu.edu.cn, nanguo2021@bupt.edu.cn)

1 Introduction

In recent years, the rapid development of wireless communication technology has profoundly reshaped various aspects of our society, driving unprecedented connectivity and enabling innovative applications [1]. Following the widespread deployment and success of the fifth-generation (5G) networks, attention has shifted towards the sixth-generation (6G) wireless communication systems. With its enhanced capabilities, it is anticipated that 6G will bring transformative changes, including ultra-low latency, significantly higher data transmission rates, increased reliability, and ubiquitous connectivity [2, 3]. Among these advancements, integrating artificial intelligence (AI) into 6G networks is expected to be a game-changer, providing new paradigms and opportunities across multiple fields [4].

AI technology has advanced rapidly over the past decade, particularly in machine learning (ML) [5], deep learning (DL) [6], and natural language processing (NLP) [7]. These advancements have enabled AI to play a crucial role across various industries. By virtue of its powerful data analysis and learning capabilities, AI can excavate massive data in wireless communication networks, realizing intelligent management and optimization of the network. In the 5G era, AI has been successfully applied to the aspects of wireless networks, e.g., network optimization, traffic prediction, and fault detection, significantly enhancing network performance and user experiences [8]. However, many unresolved issues need to be addressed to achieve AI-native support in 6G networks.

The introduction of AI can enhance coding efficiency by utilizing compressed semantic information (SI) to transmit more data with less bandwidth, which helps alleviate network congestion and improve data transmission rate [9]. However, AI algorithms generate additional data that needs to be transmitted, such as model parameters, training data, and real-time feedback. This raises the question of whether the overall data volume in future networks will decrease or increase. This change in data volume will directly impact network design and architecture [10]. Moreover, both base stations (BS) and user devices will employ AI algorithms for resource allocation to reduce energy consumption and improve resource utilization [11,12]. The operation of AI itself may increase power consumption, leading to another question: will energy consumption in future networks increase or decrease? In the pursuit of high efficiency, it is crucial to consider optimizing energy efficiency in AI applications to ensure the sustainable development of networks. Additionally, while networks can leverage AI algorithms to enhance transmission reliability and better respond to changing network conditions and user demands, the inherent uncertainty of AI algorithms — especially in complex and dynamic network environments — raises concerns about whether future networks will operate more reliably. These uncertainties may lead to decision-making errors, affecting user experience and overall network performance.

It is essential to consider integrating AI into the architecture, network elements, and functional processes of the new 6G system from a closer and deeper perspective to address these issues. Specifically, we need to re-examine the network's design philosophy to ensure a genuine synergy is formed between AI and the 6G network. This comprehensive review explores the intricate relationship between 6G and AI, delving into the fundamental principles, key technologies, and potential applications this integration brings forth. We discuss the critical technical enablers of 6G, including advanced wireless communication techniques, spectrum management, and network architectures. Additionally, we examine the role of AI in optimizing network operations, enhancing security, and enabling intelligent decision-making processes. The review also highlights the open challenges to fully realizing the potential of 6G and AI integration. This review can be a valuable resource for researchers, practitioners, and policymakers involved in developing and deploying 6G and AI technologies by providing a holistic overview of the state-of-the-art and future directions. Through this collaborative effort, we can harness the full potential of these cutting-edge technologies to build a brighter, more connected, and more sustainable future.

The organization of this paper is illustrated in Figure 1, and we summarize the main abbreviations used throughout this work in Appendix A.

1.1 Requirements and challenges in the post-5G era

1.1.1 *Better utilization of spectrum resources*

The volume of network traffic data has reached unprecedented levels [13]. Networks face the challenges of explosive growth in data traffic and the demand for massive devices to connect everything. As an essential mobile communication resource, spectrum remains the most critical element in improving network capacity [14]. Whether the limited spectrum resources can be better managed is the key to ensuring that the

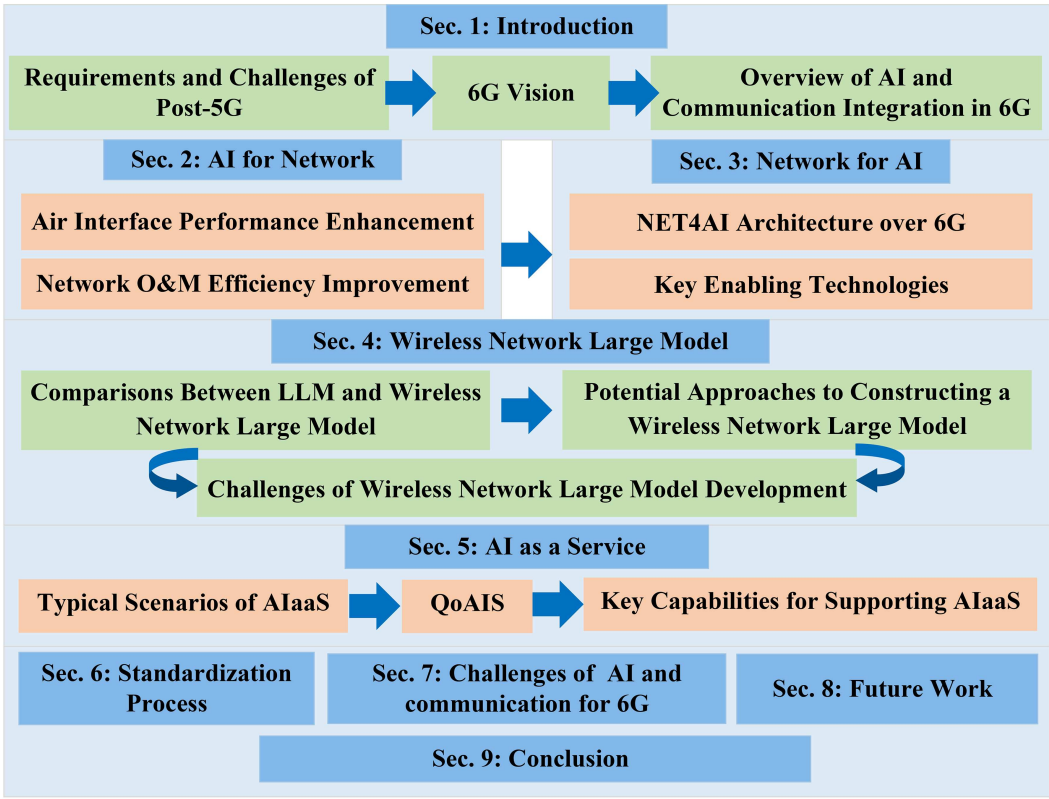

Figure 1 (Color online) Organization of this article.

network can provide high-quality services to users. However, existing wireless communication systems still adopt a static spectrum management model, where authorized users have exclusive access. This model lacks the ability of dynamic adjustment and is unable to flexibly allocate resources like spectrum and power in accordance with the real-time requirements of services. This likely leads to resource waste or insufficient resources for some services, which impacts network performance and user experiences.

1.1.2 *Lower-carbon wireless coverage*

The energy consumption of 5G systems is about three times that of the fourth-generation (4G) system due to its wider bandwidth, more channels, and more complex equipment architecture [15]. Moreover, due to the use of higher-frequency bands in 5G technology, there is a further reduction in the coverage area per BS. To achieve the same coverage targets, the number of 5G BS required would be three to four times that of 4G, significantly increasing the cost of network deployment for 5G BSs. The current operational management systems of actual networks lack flexibility and intelligence, as a BS cannot be dynamically adjusted in precise accordance with real-time changes in regional user traffic. During periods and in areas with low traffic demand, the BSs may maintain a high workload, leading to energy waste.

1.1.3 *More efficient and cost-effective network O&M*

The operation and maintenance (O&M) management of current 4G/5G networks relies primarily on manual on-site troubleshooting, resulting in low levels of automation and high maintenance costs. This approach is rather reactive, addressing issues as they arise, as opposed to proactively preventing them. Providing rapid response reports and swift emergency handling capabilities is challenging. At the same time, 5G networks introduce new technologies, such as virtualization and mobile edge computing (MEC) [16,17]. Edge sites are expanding, characterized by their large numbers, wide distribution, and heterogeneity, resulting in the expansion of maintenance teams and uncontrolled labor costs [18, 19].

Compared to conventional core network (CN) equipment's typical millisecond-level fault detection efficiency, network resource virtualization introduces new potential fault points. It is difficult for operators to delineate the responsibility boundaries among suppliers when faults occur. Using software mechanisms

for fault detection, virtual network functions at the edge site typically experience longer response times, which complicates timely network maintenance and management. Moreover, the incessantly emerging novel applications and services have presented elevated requisites with respect to network bandwidth, latency, reliability, and security.

1.1.4 *More personalized and customized on-demand service capability*

In the post-5G era, a mobile communication network, as a critical infrastructure for digital social transformation, is no longer confined to the business domain of conventional mobile communication. Instead, it places greater emphasis on the vastly diverse new demands for digital transformation across a multitude of industries. Specifically, the application scenarios of mobile communication networks are rich and diverse, covering fields such as mobile communication, mobile internet, Internet of Things (IoT), smart cities, and satellite internet. Different application scenarios have significant differences in requirements for network user experience, data rate, service latency, reliability, etc. The network needs to possess the ability to provide personalized and customized services to meet the user needs of different scenarios [20].

1.1.5 *More secure and reliable transmissions*

From 2G to 5G, the design goal of mobile communication networks in terms of security and reliability has been to ensure the authenticity of user and network identities, prevent data interception and tampering during transmission, and primarily employ security mechanisms, such as network authentication, data encryption, and integrity checks, to safeguard communications. Moreover, these security measures have undergone several "plug-in" enhancements and refinements. Establishing security measures is "post-hoc", meaning that security mechanisms are added to the system after the design of the network communication functions is complete. Furthermore, some security vulnerabilities are only addressed in subsequent generations of mobile communication networks-for example, 5G remedied the capture attacks on user identity tags in 4G [21, 22]. The defense systems are deployed primarily based on a perimeter-centric defense architecture, most often erected on the premise of "known risks". Protective devices, such as firewalls and intrusion detection systems, are positioned at the physical boundaries of mobile communication networks. Internally, these systems adopt a trusting stance, while externally, they use a "patching" approach to ward off security risks.

Additionally, the currently released 5G network security protocol standards target primarily the enhanced mobile broadband (eMBB) scenario. The standardization progress of security protocols for two other typical cases: massive machine type communications (mMTC) and ultra-reliable low latency communications (uRLLC)-lags. In the post-5G era, mobile communication networks will increasingly focus on empowering vertical industry services. They will do so by opening up specific network capabilities to third parties through upper-layer interfaces or by providing customized security services tailored to the differentiated needs of various industries [23]. This approach is intended to effectively adapt to the new characteristics of mobile communication networks, which are more open, feature heterogeneous integration, include many terminals, and exhibit diverse connection types.

1.2 6G vision

1.2.1 *Grand vision*

As new-generation information and communication technologies (ICTs) and applications such as big data, cloud computing, the IoT, and AI develop, and as these intertwine deeply with information, communication, data, and technology, the technology landscape transitions into an era of "digital twin (DT) and ubiquitous intelligence" in the 6G era [24]. 6G networks will build a closely interconnected network space by providing communication services that integrate the physical and virtual worlds, enabling seamless interaction between the human society, physical world, and virtual world. This integration will create new value through the digital world, realizing the promising vision of "6G changing the world".

In 2030 and beyond, mobile communication application scenarios will exhibit entirely new characteristics within the context of the DT world and pervasive intelligence. These scenarios will support ubiquitous wireless connectivity, big data, and new technologies such as AI, giving rise to three major application areas: intelligent living, intelligent production, and intelligent society. These areas will encompass integrated air-space-ground-sea networks, communication and sensing interconnection, and intelligent interaction. 6G networks will not be confined to providing communication functions alone.

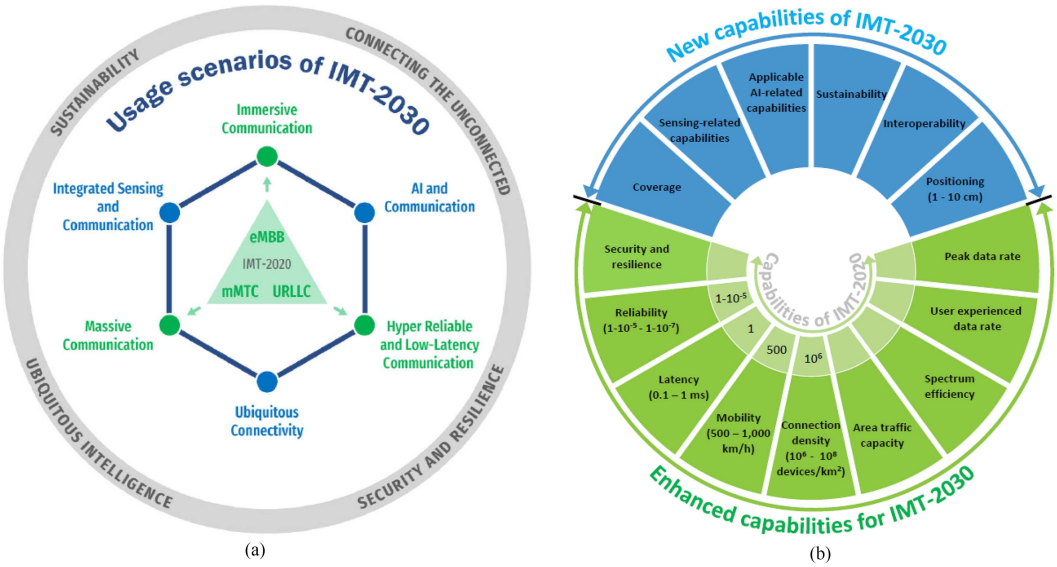

Figure 2 (Color online) Usage scenarios and new capabilities of 6G proposed by IMT-2030 [25]. (a) Usage scenarios; (b) capabilities.

Following the major trend of integrating sensing, communication, computing, AI, and security, 6G will expand from conventional single-function capabilities to new network capabilities such as sensing, computing, AI, and security.

In this diverse application context, the boundary of 6G networks has become a complex and evolving concept. Compared to conventional networks, 6G will incorporate more cross-layer designs. While retaining the functionalities of traditional networks, 6G also reflects an increasingly close integration between network operators and content service providers. 6G not only fully leverages AI technologies but also actively promotes their development. Due to these unique characteristics, the boundary of 6G cannot be simply defined as those of traditional networks. Instead, the boundary of 6G will be determined collaboratively by standardization organizations and network operators. These entities play a crucial role in establishing the technical specifications, performance requirements, and service scopes for 6G, thereby helping to clarify its boundary.

1.2.2 *Usage scenarios of 6G*

In June 2023, the International Telecommunication Union (ITU) radio sector (ITU-R) approved the "Framework and overall objectives of the future development of IMT for 2030 and beyond" [25] at the 44th Meeting of the ITU-R Working Party 5D. This document outlines the development goals, typical scenarios, and capability indicators for 6G, providing essential guidance for subsequent 6G technology and standards research. IMT-2030 (6G) has defined six significant scenarios. As shown in Figure 2, building on the "iron triangle" of IMT-2020 (5G), IMT-2030 (6G) extends outward to form a hexagon. On the outermost circle of the hexagon, four design principles applicable to all scenarios are listed: sustainability, ubiquitous intelligence, security/privacy/resilience, and connecting the unconnected.

• **Immersive communication.** The immersive communication scenario extends the eMBB of 5G. It includes use cases that provide users with rich interactive video (immersive) experiences, including interactions with machine interfaces. Typical use cases in this scenario include immersive extensive reality (XR) communication, remote multi-sensory presentation, and holographic communication. In immersive communication, supporting time-synchronized mixed traffic of video, audio, and other environmental data is essential, including independent support for voice. Moreover, the ability to improve spectrum efficiency, provide a consistent service experience, and balance higher data rates and enhanced mobility in various environments is crucial. Some immersive communication use cases may also require high reliability and low latency support to enable responsive and accurate interactions with real and virtual objects and greater system capacity to connect numerous devices simultaneously.

• **Hyper-reliable and low-latency communication.** The hyper-reliable and low-latency communication scenario extends the uRLLC capabilities of 5G, covering specialized use cases that are expected

to have more stringent requirements for reliability and latency. Typical use cases include communications for comprehensive automation, control, and operations in industrial environments, such as robotic interactions, emergency services, remote medical care, and power transmission and distribution monitoring.

• **Massive communication.** The massive communication scenario expands upon the mMTC capabilities of 5G, focusing on connecting many devices or sensors. Typical use cases include smart cities, transportation, logistics, healthcare, energy, environmental monitoring, agriculture, and many other fields with expanded and new applications. This scenario involves various IoT devices that may not have batteries or have long-life batteries. This scenario requires support for high connection density. Depending on the use case, the scenario requires different data rates, low power consumption, mobility, extended coverage range, and high security and reliability.

• **Ubiquitous connectivity.** The ubiquitous connectivity scenario aims to enhance connectivity to bridge the digital division. One of the critical focuses of this use case is to address areas that currently lack coverage or have minimal coverage, especially in rural, remote, and sparsely populated areas [26,27]. Typical use cases include, but are not limited to, IoT and mobile broadband communication in these underserved regions.

• **AI and communication.** The AI and communication integration scenario will support distributed computing and AI-driven applications. It will enable unprecedented use cases by leveraging data collection, local or distributed computation offloading, and distributed training and inference of AI models across intelligent nodes. Typical use cases include assisted autonomous driving, autonomous collaboration between medical devices, offloading intensive computations across devices and networks, creating and predicting DT, and assisting collaborative robots. This scenario requires support for high regional traffic capacity, user experience data rates, low latency, and high reliability tailored to specific use cases. In addition to communication aspects, this usage scenario is expected to include a range of new functionalities integrating AI and computing capabilities into IMT-2030. These include data collection, preparation, and processing from diverse sources, distributed AI model training, model sharing, and distributed inference across IMT systems, as well as coordination and linking of computing resources.

• **Integrated sensing and communication (ISAC).** This scenario contributes to new applications and services that require sensing capabilities. It leverages IMT-2030 to provide wide-area multidimensional sensing, offering spatial information about unconnected devices and spatial information about connected devices, movement, and the surrounding environment. Typical use cases include IMT-2030 assisted navigation, activity detection, and motion tracking (e.g., posture/gesture recognition, fall detection, vehicle/pedestrian detection), environmental monitoring (e.g., rainwater/pollution detection), and providing sensory data/information about the surrounding environment for AI, XR, and DT applications. Besides the provided communication functions, this scenario also requires support for high-precision positioning and related sensing capabilities, including distance/velocity/angle estimation, object and presence detection, localization, imaging, and mapping.

1.2.3 *Capabilities of 6G*

Regarding development goals, 6G aims to achieve seven significant objectives: inclusivity, ubiquitous connectivity, sustainability, innovation, security/privacy/resilience, standardization and interoperability, and accessibility. It may become a new digital infrastructure that better connects the physical and virtual worlds, supports new users, and empowers new applications. Regarding performance indicators, the Recommendation [25] specifies 15 key capability metrics for 6G, divided into two categories. The first category focuses on enhanced capabilities for IMT-2020, including peak rate, user experience rate, spectrum efficiency, regional traffic density, connection density, mobility, latency, reliability, and security/privacy/resilience performance-totaling nine indicators. The second category supports new functionalities for extending IMT-2030 use cases, encompassing coverage, sensing, AI, sustainability, interoperability, and positioning-totaling six indicators. Each capability may exhibit varying relevance and applicability across different usage scenarios. The ITU's vision for 6G encompasses a range of services and scenarios to advance communication capabilities beyond what the current 5G achieves.

Among the above-mentioned newly added capabilities, the applicable AI-related capabilities have drawn the most attention. The possession of AI-related capabilities by 6G networks can be understood from several aspects. Firstly, the network can optimize its own performance through AI. Secondly, the network is capable of providing support for the operation of AI. Eventually, the network can offer AI services to users and equipment just as it provides communication connections. Such capabilities can

Table 1 Data information.

Data type	Data name	Application direction
Air interface data	Channel state information, intra/inter-cell interference, multipath delay	Channel state information prediction, wireless channel modeling, power control, interference cancellation
Terminal data	Measurement reports, minimization of drive tests, block error rate, port flow	Network traffic detection and congestion control, traffic prediction and scheduling optimization, personalized service
Network data	KPI, extended detection and response, running log	Dynamic load balancing and interference avoidance in wireless networks, network energy conservation, intelligent routing
Business data	IP information, uniform resource locator information, subscriber tariff data, internet access time	Business identification and awareness, business anomaly detection, traffic management

only be achieved through the in-depth fusion of wireless communication and AI.

1.3 Overview of AI and communication integration in 6G

This subsection elaborates on why and how to integrate AI and wireless communication. Generally speaking, AI can endow wireless networks with "autonomy" and "intelligence", while wireless networks provide AI with a broad application space. These two inseparable strategic development fields will continuously promote the convergent innovation and technological innovation of information, communication, and computing, and realize people's visionary prospects in the spatiotemporal domain, information interaction types, and cross-industry convergent and innovative applications.

1.3.1 *Drivers of integrating AI into wireless communication*

Future wireless network systems will evolve into integrated wireless infrastructure platforms deeply integrating "communication", "sensing", "computing", "intelligence", and "storage". These platforms will be capable of providing customized or personalized services on demand, surpassing the capabilities of conventional communication systems that rely on fixed architectures and predefined rules. Such systems will need to leverage AI technologies. They will utilize big data and knowledge bases for inference and decision-making, employing models' generalization abilities to adapt to various environments and scenarios and providing optimal resource allocation and management solutions. This subsection analyzes the driving forces of the integration of wireless communication and AI from the following two perspectives.

(1) AI to push limits of wireless communication systems. AI possesses remarkable capabilities in handling big data. Mobile communication networks generate vast and diverse data every moment [28, 29]. The international data corporation has predicted that by 2025, the global daily data generation will reach 491 exabytes [30]. These data typically include terminal, wireless air interface, network, and service data in various formats, such as text, images, extensible markup language, hyperText markup language, graphics, and audio/video information, as shown in Table 1. Network operators can use AI to train and make inference decisions on different types of mobile big data, enhancing network performance and optimizing across various dimensions and objectives.

AI has predictive capabilities absent in conventional algorithms. AI can learn from historical data via ML algorithms and construct models to forecast future network requirements and possible issues, including network congestion, device failures, or alterations in user behavior. For instance, AI can anticipate potential traffic peaks during specific holidays or large-scale events and take resource allocation measures in advance, such as increasing the capacity of temporary BS or optimizing routing strategies, thereby effectively averting network breakdowns and enhancing user experience.

Adaptivity is a prominent characteristic of AI. The wireless communication environment is highly dynamic and affected by various factors, such as weather, terrain, buildings, and changes in the location of mobile devices. The adaptive ability of AI enables it to make rapid adjustments according to environmental changes and fluctuations in network conditions. For example, when a user moves from indoors to outdoors or is in a high-speed vehicle, the signal strength and interference situation will change. AI can

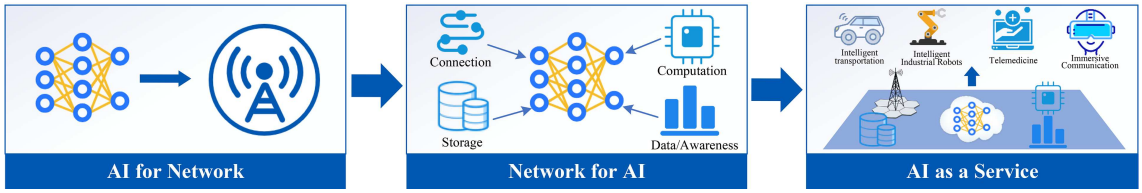

Figure 3 (Color online) Extent of 6G and AI integration: AI for network, network for AI, and AI as a service.

sense these changes in real time and automatically adjust communication parameters to ensure the stability of communication quality, adapt to different communication scenarios, and improve the reliability and flexibility of wireless communication systems.

In contrast to the hierarchical management of different levels in conventional communication systems, in theory, AI can learn hidden structures and parameters to fit arbitrarily complex functions. This provides a more effective way to sense the complex and variable wireless environment and characterize the network state space. On the other hand, the designs of different communication system modules might have conflicting goals, and there might be performance constraints between modules. There are trade-offs between performance metrics, such as channel capacity and interference, transmission reliability, and system energy consumption. Optimizing each module individually often fails to achieve overall optimal performance. In such cases, AI can facilitate joint optimization design among modules.

(2) AI for resolving challenges in wireless communications. The core reason for the challenges summarized in Subsection 1.1 is that the current wireless network is not intelligent enough. Future wireless network systems will evolve into an integrated wireless infrastructure platform that deeply integrates "communication", "sensing", "computation", "intelligence", and "storage".

On the wireless radio access network (RAN) side, AI can be combined to conduct data analysis and decision-making for the control unit plane of 5G new radio (NR), realizing the immediate dynamic performance optimization of wireless resource scheduling and management. The next generation node B (gNB) BS can achieve intelligent on/off handoff based on AI to achieve dynamic energy savings. On the CN side, with the network data analytics function (NWDAF) network elements achieving comprehensive interconnection with surrounding CN elements and realizing data collection, AI will also participate in the control and decision-making of the CN, including quality of service (QoS) guarantee, traffic handing, 5G edge computing, and load balancing of network functions. International communication standard organizations, such as ITU, 3rd Generation Partnership Project (3GPP), and IMT-2030, have initiated the standardization of AI applications in communication systems, and made significant progress. In the future, AI is anticipated to further empower the infrastructure within networks, network management, operation systems, and service systems, fully unleashing the potential of integrating wireless communication and AI in the communication ecosystem and applications.

1.3.2 *Vision of AI and 6G integration*

The advancement of AI technology has significantly empowered the development of wireless networks. Figure 3 illustrates the degree of integration between 6G and AI from three perspectives: AI for network (AI4NET), network for AI (NET4AI), and AI as a service (AIaaS). The application of AI to enhance wireless networks is generally referred to as AI4NET. Its essence lies in using AI to improve wireless networks' performance, efficiency, and user experience. AI4NET's application in 5G networks has significantly facilitated the intelligent development of mobile communication networks and vertical industries. Its role primarily manifests in O&M intelligence and network element intelligence [31]. The former emphasizes utilizing AI to optimize conventional algorithms and automate and intellectualize tasks across network layers and operations. The latter focuses on proposing AI-driven intelligence for individual network elements and functional entities, breaking down proactive network responses into predictions of multiple performance metrics. This approach ultimately aims to optimize network resource allocation and reduce network latency.

To meet the transformative and developmental demands of AI in wireless networks, 6G will no longer be merely a communication network that serves as a connectivity enabler for intelligent services. Instead, it will evolve into an integrated information network that combines communication, sensing, and computing functionalities. While AI empowers 6G networks, 6G networks will also empower AI, achieving NET4AI, which provides comprehensive support for the deployment and application of AI in wireless environ-

ments [32, 33]. The essence of NET4AI is to provide AI with various support capabilities, enabling more efficient and real-time AI training/inference and enhancing data security and privacy protection. The IMT-2030 (6G) Promotion Group in China envisioned the 6G network architecture in [34], as an open and innovative platform for information services, offering capabilities that transcend mere connectivity. These capabilities encompass computing power networks, trusted security, sensing, and data services, which are essential for the operation of AI in networks. The authors of [35] proposed a novel network architecture for providing native support for AI, called an AI-oriented network, which is represented as a network management framework with distributed AI computational capabilities and multi-party participation built in 6G networks.

Moreover, to effectively support "inherent intelligence" and achieve "native AI", 6G networks will treat AIaaS for provision and processing. This concept gives rise to the idea of 6G AIaaS. Specifically, 6G AIaaS utilizes resources and functionalities within the network (including 6G CN, wireless access networks, and terminals), such as connectivity, computing, data, and models. It aims to construct a distributed, efficient, energy-efficient, and secure AI service ecosystem, which includes AI model training, inference, deployment, and other functionalities in a low-carbon open environment [36]. Not only can it redefine the ecosystem of edge cloud, but also build new business models through 6G mobile networks, transitioning from past connectivity-oriented networks to service-oriented networks, ultimately achieving ubiquitous intelligence [37, 38]. AIaaS's typical scenarios include but are not limited to, smart cities, smart agriculture, universal education, and smart industry. Typical applications include unmanned taxi services, smart grid inspections, home health monitoring, and virtual classrooms.

Large language models (LLMs) [39] have represented a significant breakthrough in NLP and can potentially contribute to the development of 6G AI. Compared to traditional smaller parameter models, LLMs exhibit strong contextual understanding, coherent text generation, logical reasoning, and generalization capabilities. Existing general-purpose large models, such as the closed-source GPT model developed by OpenAI and the open-source Llama model developed by Meta (formerly Facebook), can address a wide range of domain-specific problems [40, 41]. Recently, the open-source DeepSeek LLM developed by the DeepSeek team in China has potentially become a game-changer, demonstrating the feasibility of training large-scale AI models (LAMs) at unprecedentedly low cost [42]. LLMs can potentially play a critical role in communication systems, e.g., in task processing and intelligent services.

The emergence of LLMs can bring a new paradigm for the integration of 6G and AI. The intrinsic AI architecture in 6G can provide linkage, computation, model decomposition, and model distribution services for the training and inference of LLMs. LLMs can empower various domains of 6G networks (e.g., air interface, network side, network security, network O&M, etc.), enabling intelligent services and closed-loop control. In 6G networks, customized specialized large models can be deployed across different network layers. These specialized large models can collaboratively address network issues through task decomposition and composition, cross-layer collaboration, and cloud-edge collaboration, enhancing the intelligence level of the network.

In addition to using customized AI models tailored for specific functionalities and tasks, large models demonstrate superior generalization performance across multiple tasks and breakthrough capabilities on complex tasks that small models cannot achieve. The functionalities of large models increasingly align with the multi-scenario, multi-task characteristics of 6G, effectively alleviating the workload in model perception and performance metric setting within 6G networks and showing broad application prospects. A possible form of integrating 6G with AI is illustrated in Figure 4. Leveraging data from diverse heterogeneous networks within 6G, these large models can synthesize different 6G scenarios and services, laying a crucial foundation for AI empowerment in 6G networks. For example, the proposal of 6GANA in the 6G NETGPT white paper has built an LLM similar to ChatGPT, which provides a new paradigm for advanced operations and management of 6G networks [38].

2 AI for network

The most fundamental goal of integrating wireless communication and AI lies in enhancing the performance, and efficiency of the network and the service experience of users by virtue of AI, which is known as AI4NET [43]. At this stage, the focus is on how to utilize AI algorithms to optimize communication performance and network functions. For example, AI can optimize the signal modulation and demodulation process, making signal transmission more accurate and efficient; with the help of AI, network

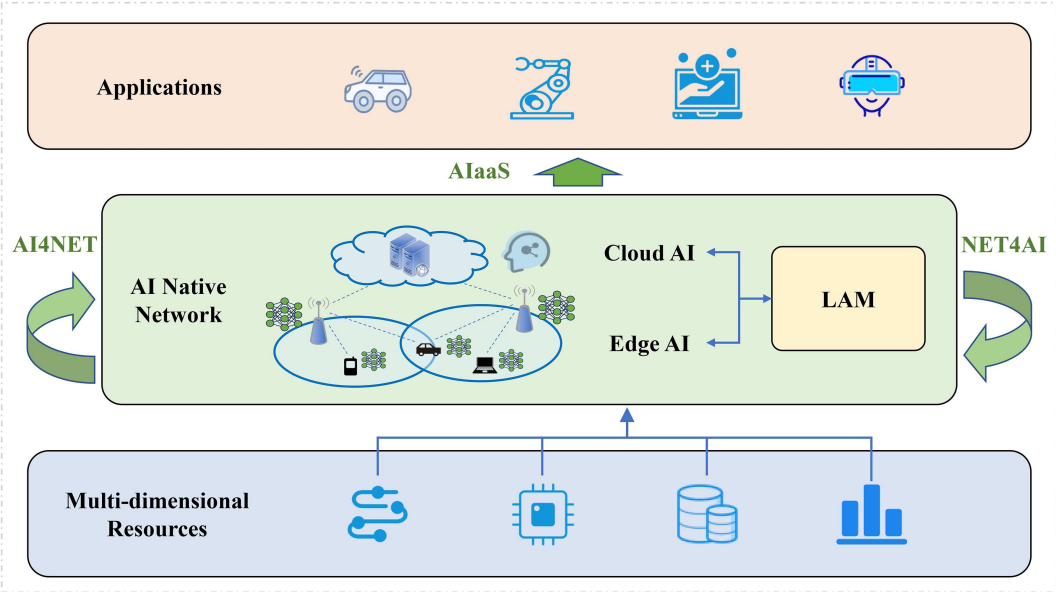

Figure 4 (Color online) Design of 6G AI integration.

resources can be intelligently allocated to achieve load balancing; and an automated O&M management mode can be constructed to improve O&M efficiency and reduce costs. This introduction of AI is not anticipated to have a significant impact on the original network architecture. Instead, it is to improve the communication problems in specific recognition by training AI algorithm models. This process is analogous to precisely implanting intelligent patches in the network. It can not only maintain the stability of the network architecture but also gradually enhance the intelligence level of the network.

The concept of AI4NET is presented in Figure 5, where the wireless AI possesses capabilities such as feature extraction, prediction, adaptation, optimization, real-time processing, correlation, and scene clustering. These capabilities form the foundation for AI-enabled wireless networks and can be directly deployed on the BS side and the CN to empower wireless communication. AI optimizes conventional methods to enhance network performance on various terminal devices. On the network side, AI takes advantage of its capabilities to improve the quality of end-to-end services, optimize the functions of network elements, conduct automated management of the upper-layer network, and boost network O&M efficiency. The advent of AI has also given rise to a series of new services aimed at improving and serving users in a better manner.

AI4NET has initiated relevant research and applications in 5G. In 6G, with the more mature AI technologies represented by DL and the emergence of new infrastructures integrating connectivity and computing power, the relevant applications will become more abundant and sophisticated and may further evolve in depth. This section elaborates on some cases from two aspects, namely air interface performance enhancement and network O&M improvement, to illustrate the concept of AI4NET.

2.1 Air interface performance enhancement

In this subsection, we show how AI technology can improve transmission in channel state information (CSI) feedback, orthogonal frequency division multiplexing (OFDM) receivers, beam management, and wireless localization.

2.1.1 *AI-based CSI feedback algorithm*

Within the 6G framework, the feedback overhead is expected to experience a sharp increase along with a significant rise in the number of antennas. Consequently, ensuring the accuracy of channel reconstruction and minimizing the CSI feedback cost is one of the bottleneck problems that need to be surmounted. The utilization of AI technology in CSI feedback can profoundly dissect these elaborate channel characteristics. Through training with a substantial quantity of labeled samples, the AI model can acquire the characteristic patterns of the channel within diverse fading environments, interference scenarios, and multipath circumstances. In contrast to conventional CSI estimation approaches relying on statistical

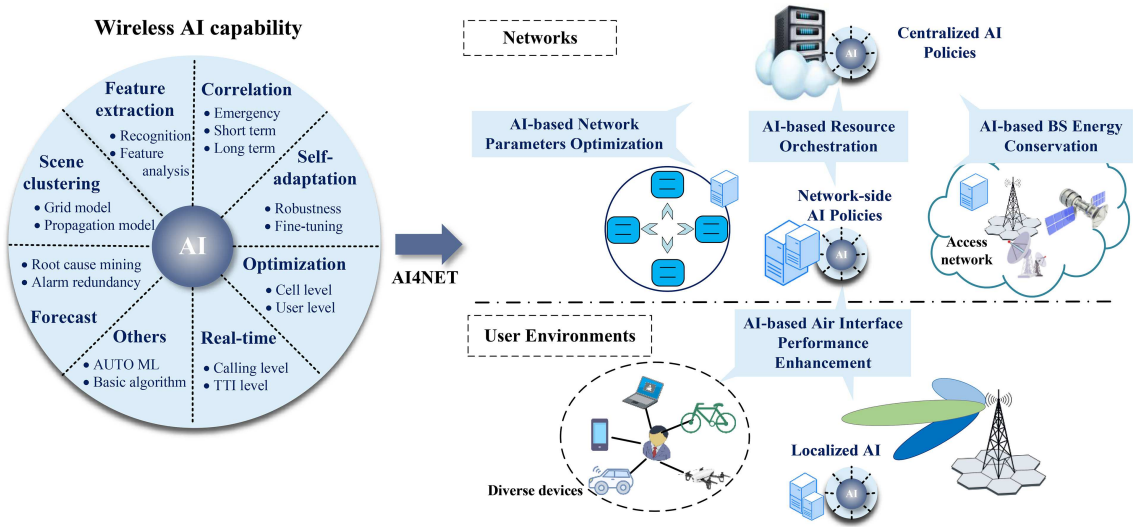

Figure 5 (Color online) Schematic of AI's capabilities for network.

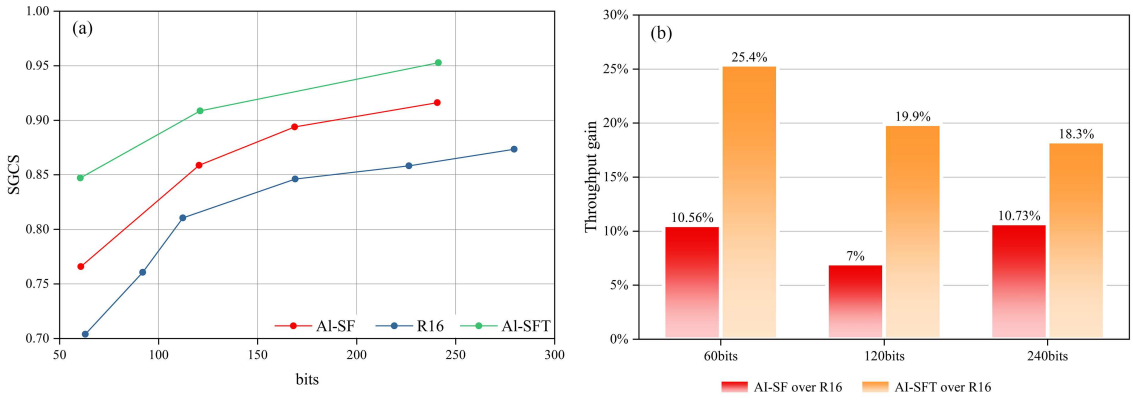

Figure 6 (Color online) Performance of R16 Type II codebook, AI-SF, and AI-SFT on SGCS and throughput gain. AI-SF means the AI/ML-based spatial-frequency compression method, which transformer is used as the backbone of both the encoder and decoder. AI-SFT is AI-SF with additional past CSI. (a) SGCS versus bits; (b) throughput gain versus bits.

theory [44] or simple linear models [45], AI-based methods exhibit stronger generalization competencies and can estimate CSI with greater precision in complex real-world communication settings.

The massive CSI data and the inherent random characteristics of CSI make AI potentially useful for designing a new CSI feedback mechanism. For instance, the nonlinear characteristics of DL can be utilized to efficiently extract the features of CSI, and the original CSI data can be transformed into a more compact feature representation, thereby reducing the communication overhead in the CSI feedback process. An example of its implementation is the autoencoder architecture, which employs neural networks to extract and compress the features of CSI. The authors of [46] were the first to apply DL to CSI feedback. The proposed CsiNet scheme utilizes convolution to extract channel features, and compress and reconstruct channels, and it is superior to the CSI feedback scheme based on compressed sensing in terms of feedback accuracy and computational complexity. The encoder of CsiNet compiles the high-dimensional CSI into code words, while the decoder is responsible for decoding the codewords back into the original CSI. The internal structures of encoders and decoders of different algorithms vary, resulting in differences in channel reconstruction performance. As shown in Figure 6, Huawei demonstrated the huge advantages of AI-based CSI feedback over conventional codebook approaches through system-level simulations. AI with additional information can effectively improve the performance on the square of the generalized cosine similarity (SGCS) and throughput [47].

The study in [48] increased the size of the convolutional kernel based on CsiNet to improve the perceptual field of view of the convolutional layer, which is conducive to extracting the sparsity of the channel and enhancing the accuracy of channel reconstruction. Although a larger perceptive field of view could

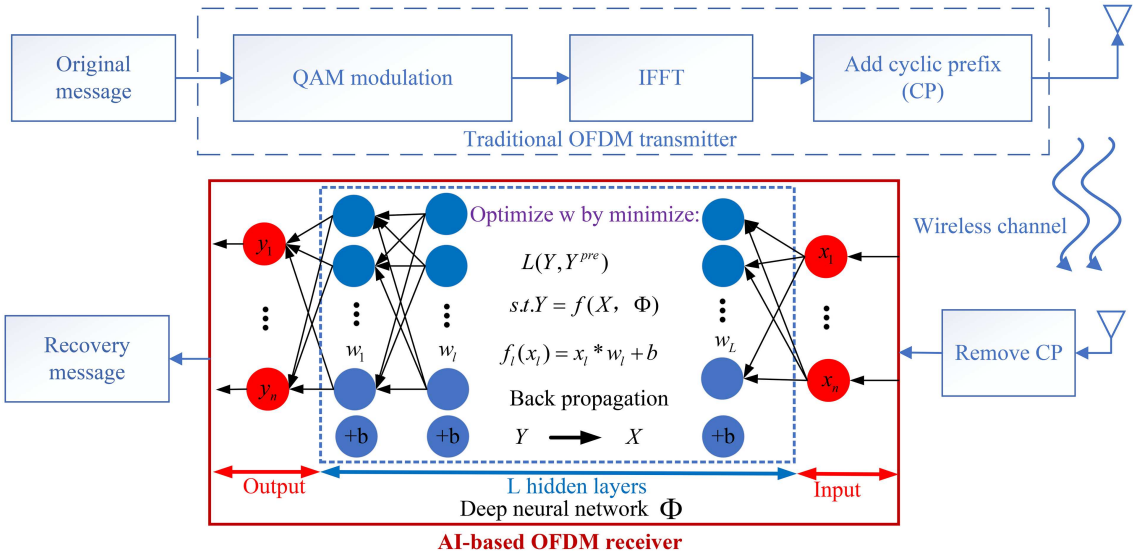

Figure 7 (Color online) AI-based OFDM receivers. Using AI modules to replace conventional physical layer processing, AI's capabilities can compensate for signal distortion.

effectively extract the sparsity of CSI, a smaller convolution kernel could extract finer features from CSI. A DL-based CSI compression and quantization method was developed in [49] considering high compression ratios. A multi-rate CSI compression framework was designed in [50] to improve the generalization of the model in the field of transfer learning. In [51], a non-local neural network was introduced based on the CsiNet network to capture a wide range of dependencies, and the accuracy of channel recovery was improved compared to CsiNet. The CRNet was proposed in [52], which used convolution kernels of different sizes for channel feature extraction and recovery, reducing the computation and improving the accuracy of channel reconstruction. In order to better deploy AI-based CSI feedback models in real-world wireless environments, model monitoring, model updating and AI-based signaling process management are hot issues that require our attention in future studies.

2.1.2 *AI-based OFDM receiver*

AI method can extract essential information in sparse and time-varying pilot frequencies. The complex mapping relationship between the input and the output is effectively constructed, thus AI-based receiver can learn and adapt to channel characteristics in complex environments. Each module of the conventional receiver requires accurate signal modeling and calculation, leading to the complexity and calculation burden of the system. AI has the potential to extract features directly from the original signal and demodulate it in a data-driven manner, simplifying the complex process of conventional receivers.

Regarding wireless OFDM receivers, some researchers have used a fully connected deep neural network (DNN) to improve the existing modular OFDM receiver. A typical framework of an AI-based receiver is given in Figure 7. We use DNN parameters Φ to construct the relationship between the input X and output Y as $Y : f(X, \Phi)$. With the assumption of L hidden layers, we denote w_l as the parameter weights of the l-th layer, where l ranges from 0 to $L + 1$. w_0 and w_{L+1} represent the weights of the input layer and the output layer, respectively. b denotes the bias of each layer. In the training stage, the optimization direction of the model minimizes the loss between the output Y and true label Y^{true}. Thus, the network parameters Φ are updated by changing w_l during the process of back propagation. Hence, the AI model can encode the information of actual scenes into neural networks and optimize the performance of algorithms by adjusting the structure and parameters of neural networks.

In [53], a model-driven DL approach was proposed, which combines DL with expert knowledge to replace existing OFDM receivers in wireless communications. An ML-assisted physical layer receiver technique was proposed to demodulate the OFDM signals, subject to very high Doppler effects and corresponding distortions in the received signal [54]. To strike a balance between full-size cyclic prefix (CP) and non-existent CP, the authors of [55] investigated the redundancy problem and proposed a minimum redundant OFDM receiver using DL tools. In [56], the receivers were designed based on DNN, consisting of a layer of DNNs and soft decisions. The problems of channel estimation error, delay, and

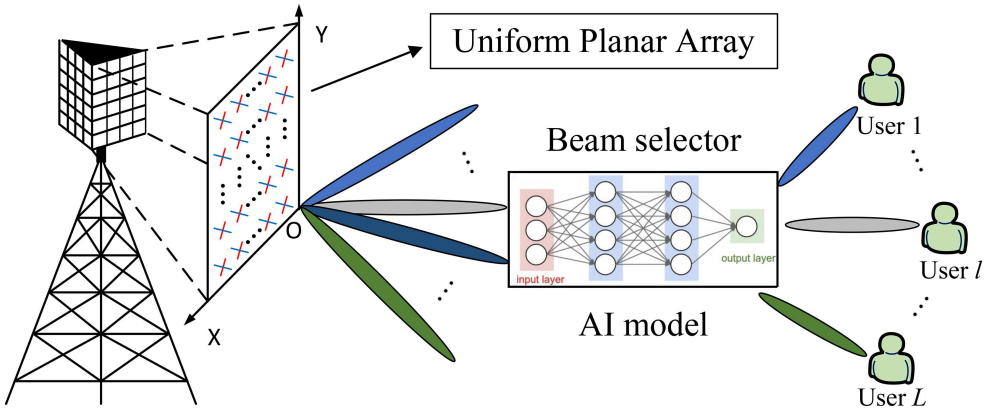

Figure 8 (Color online) Structure diagram of AI-based beam management. AI selects the optimal beam by looking for the complex mapping relationship between input and output, which can significantly reduce the search cost of the beam while ensuring prediction accuracy.

decoding limits between users with conventional detection methods were solved. The symbols of all users were recovered at one go to jointly perform channel estimation and signal detection. A novel generative supervised DNN was designed in [57], which could be trained using a reasonable number of pilots. After channel estimation, the neural network-based receiver jointly learns the pre-encoder and decoder for data symbol detection. An intelligent receiver for OFDM communication systems was designed in [58] based on the DNN structure, and realized by optimizing the DNN structure. This method can recover information on the receiving side and avoid complicated pilot operations and signal error accumulation. Moreover, a convolutional neural network (CNN) was used to reduce the bit error rate using the mathematical calculation function of discrete Fourier transform and the training of OFDM signal samples [59].

2.1.3 *AI-based beam management*

In the existing literature, the beam selection methods of multi-antenna systems can be divided into beam selection based on beam scanning [60], beam selection based on position prediction [61], and beam selection based on hierarchical search [62]. Conventional beam management methods have some limitations when applied to millimeter wave massive multiple-input multiple-output (MIMO) systems [63]. The primary cause is that the real-world wireless propagation environment often has various characteristics. After a beam is narrowed, the signal transmission is more susceptible to occlusion. The increase in the number of beams significantly increases the number of beam search operations in beam alignment. It is difficult to accurately select the optimal beam instantaneously.

With the application of AI, the optimal beam can be selected without scanning all beam pairs, thus alleviating the scanning overhead existing in the conventional beam management method, as shown in Figure 8. In [64], a DNN was utilized for beam alignment based on the contextual information of the UE location. This method can reduce the search time by four times for AP selection and over tenfold for beam selection. By taking advantage of the mmWave channel structure, a novel hierarchical beam alignment framework was proposed in [65] to leverage DL techniques to seek the optimal beamformer with learnable probing codebooks. This novel framework learns two tiers of probing codebooks and uses their measurements to predict the optimal beam in a coarse-to-fine search manner. A DL-based beam alignment method was proposed in [66] by jointly training the probing codebook and the beam predictor. Simulation results demonstrate that the method can achieve high beam alignment accuracy while reducing the beam sweeping complexity by ten times.

The above literature aims to use DL methods to improve the accuracy and robustness of beam alignment, and has demonstrated the superiority of AI. For practical deployment, we should further consider the acquisition of high-quality tags in complex environments. Considering the training overhead, improving the generalization ability of AI models among different BS environments also becomes an important future work.

2.1.4 *AI-based wireless localization*

The 6G network is mandated to furnish high-precision wireless location capabilities for diverse application scenarios. This location information is usually in the form of the mobile user's geographic coordinates relative to some reference points [67]. Many applications need high-precision positioning, such as industrial automated guided vehicle and asset tracking, especially indoor precision positioning, but the global positioning system (GPS) cannot be used indoors. Despite having a specific positioning reference signal, LTE cannot meet the requirements for high-precision positioning due to its relatively low positioning accuracy and the large distance between BS (about 100 m even with a 20 MHz bandwidth). Bluetooth, Wi-Fi, and other wireless location-based technologies can have high deployment costs and are difficult to become a universal positioning technology [68].

The application of AI technology has been introduced to achieve high-precision positioning in the 5G era. One approach is to utilize the random forest technique [69], which can create a classification model to split a vast area into many small grids and then predict the grid where the user is located. It can also make a regression model to predict the user's position coordinates. In another method, multi-layer perceptron has a similar application to random forest and can establish regression and classification models to locate targets [70]. AI algorithms can also be trained to recognize the fingerprint of a wireless signal in a specific location and fuse data from different sensors to achieve high-precision positioning.

With the widespread use of massive MIMO technology, AI has shown more significant advantages in high-precision positioning. Using the location in the line-of-sight (LoS) and non-line-of-sight (NLoS) channel co-existence scenario based on the neural network as an example, ML and a significant volume of channel data can help effectively map the relationship between channel response [71] and position coordinates, resolving the location problem in complex environments and increasing location accuracy. Very high positioning accuracy is achieved without significant additional cost, even about 20 mm accuracy under indoor LoS and NLoS conditions. Due to the diversity of data modes in complex environments, the further development and practical application of AI-based localization requires us to consider the robustness and reliability of the AI model to enable mobility management.

2.2 Network O&M efficiency improvement

AI can analyze a large amount of data and make real-time decisions, which is especially useful in wireless networks where many variables and parameters must be continuously monitored and adjusted to optimize system performance [72]. It can dynamically allocate resources such as bandwidth and power in wireless networks to optimize system performance. In addition, the integration of AI enables the network to better understand and predict the complexity and dynamics of the network. Based on the prediction results, proactive and timely network maintenance can be carried out. By leveraging AI's DL and pattern recognition technologies, a more in-depth understanding of network behavior can be achieved, realizing more optimal resource management and scheduling strategies. Subsequently, the role of AI in network O&M management will be introduced through specific cases.

2.2.1 *AI-based traffic prediction*

The progressive deployment of 6G networks [73] heralds the increasing prevalence of new application scenarios like virtual reality (VR) and immersive communication. In such applications, the network traffic exhibits high variability. By predicting network traffic, it is possible to identify traffic fluctuations and peak periods in advance, thereby enabling dynamic allocation and scheduling of network resources. For instance, the authors of [74] employed a recurrent neural network (RNN) to achieve joint spatiotemporal prediction, aiming to improve the accurate modeling of network traffic variations. The method proposed in [75], based on deep belief networks, effectively predicts the long-term variation trends of network traffic. The result in [76] demonstrated that, compared to conventional autoregressive integrated moving average models, long short-term memory outperforms in traffic prediction tasks, particularly in handling nonlinear and complex time-series data.

In centralized traffic prediction methods, a BS typically needs to collect traffic data from various geographical locations, which inevitably introduces additional communication overhead and potential security risks. As a result, federated learning (FL)-based traffic prediction methods have emerged. The FedDA method was proposed in [77], which effectively reduces the delay and bandwidth overhead associated with data transmission by only transmitting model parameters, instead of raw traffic data. This

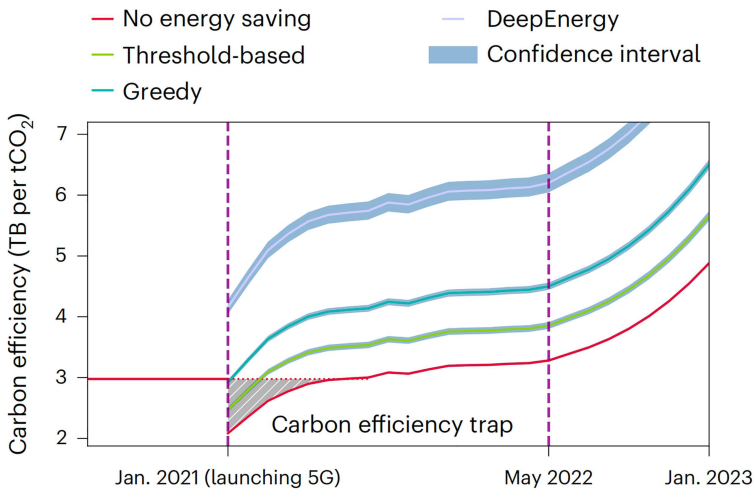

Figure 9 (Color online) Carbon emissions of different strategies [88].

approach not only ensures privacy but also improves the efficiency of data processing. The authors of [78] proposed a distance-weighted federated traffic prediction method that better captures the spatiotemporal characteristics of traffic and improves the accuracy of traffic forecasting.

To reduce the model training time, the authors of [79] proposed a transfer learning-based approach, which accelerates the training process by uncovering the potential correlations between traffic patterns in different regions. Other studies have employed techniques such as data sharing [80] and meta-learning [81], to further optimize the performance of traffic prediction models. In the future, network traffic prediction methods will extract spatiotemporal features and long-term dependencies, but also enhance the lightweight design and privacy protection capabilities of the models.

2.2.2 *AI-based BS energy conservation*

The ICT industry directly contributes approximately 4% of global greenhouse gas emissions, with mobile communication networks accounting for over 10% of this figure [82]. In response to the challenges posed by climate change, the ITU released Recommendation 1470, titled "Greenhouse gas emissions trajectories for the information and communication technology sector compatible with the UNFCCC Paris Agreement", which calls for mobile operators to achieve a 45% reduction in carbon emissions from 2020 to 2030, aiming to drive the industry's transition towards sustainability.

The energy optimization within 6G networks exhibits a close correlation with user latency. A key focus of academic research has been how to achieve network energy savings by effectively shutting down BS while ensuring the maintenance of basic service quality. The authors of [83] proposed a BS shutdown strategy based on traffic thresholds. During peak network periods, this approach may result in frequent state handoffs, leading to service interruptions. The authors of [84] classified BS states into four different sleep modes based on the activation time scale of the BS. While deeper sleep modes can significantly reduce energy consumption, they also result in increased wake-up delays, affecting service quality.

The energy-saving strategy for BS predicated on reinforcement learning (RL), via the incessant interactive learning between the agent and the operational milieu, dynamically modifies the parameters to accommodate the variations in service load and the oscillations in channel conditions. In [85], a BS state management scheme was proposed based on Q-learning. It dynamically adjusts the energy consumption of the BS in response to load changes by exploring the optimal duration of different sleep modes. The authors of [86] considered the interference between multiple BS, and designed an energy optimization method based on collaborative strategies. By enabling cooperation between BS, this method reduces the overall system energy consumption. A cascading RL method was proposed in [87], which optimizes the decision process of user management and BS states in multi-BS scenarios, further improving network efficiency.

Introducing renewable energy, especially clean energy sources, e.g., solar power, as the energy supply for networks can significantly improve network energy efficiency and reduce dependence on conventional power grids. As depicted in Figure 9, the authors of [88] integrated RL with graph neural networks

Table 2 Various parameters in the current networks.

Parameter type	Parameter name	Application area
Terminal parameter	Block error rate	Scheduling optimization
	Traffic information	Traffic monitoring
	Measurement report	Adaptive coding
Air interface parameter	Multi-path time delay	Channel modeling
	Inter-cell interference	Interference cancellation
	Channel state information	Channel state prediction
CN parameter	Operation log	Fault handling
	Network topology	Load balancing
	Network energy consumption	Energy saving
Service parameter	Online duration	Personalized service
	Transport protocol	Transmission control
	Consumption history	Service monitoring

and accomplished the optimization of energy conservation and carbon emissions in wireless networks by capitalizing on photovoltaic solar technology. Energy-saving efforts in wireless networks should be deeply integrated with the use of renewable energy, providing more sustainable solutions.

2.2.3 *AI-based network parameter optimization*

The rational setting of network parameters is crucial, as these parameters determine the design, energy consumption, and performance of the 6G network, and must be aligned with the real-time distribution of users and the electromagnetic environment. Wireless network parameters can be categorized into two types: BS parameters and threshold parameters. The BS parameters involve basic handoff functions such as paging and slicing. The threshold parameters configure the thresholds for cell reselection, event decision-making, and neighbor cell handoff. Moreover, spectrum resource planning in communication systems is a critical component of parameter optimization. Studies have shown that spectrum is often underutilized [89], and AI technologies have demonstrated great potential in cognitive radio [90]. AI can monitor and mitigate interference [91], predict spectrum usage trends [92], and significantly improve the utilization of spectrum resources.

Conventional parameter optimization relies mainly on drive tests and expert experience. When network operators identify significant deficiencies in communication service quality, specific operators will deploy drive test vehicles to collect and report network performance metrics, such as network throughput and audio/video transmission delay. Subsequently, operators create scatter plots to further characterize network performance. Finally, expert experience is used to analyze and optimize the network parameters [93]. Due to the significant increase in the number of 6G antennas, the channel is not a simple one-dimensional circuit loss structure but has expanded into an N-dimensional angle power spectrum. The efficiency of expert-based parameter tuning is low, making it difficult to modify billions of 6G network parameters. Furthermore, conventional experience-based tuning methods often rely on inherent rules, which makes it difficult to fully capture and adapt to the complex relationships in the network [94]. Table 2 summarizes various parameters in the current networks.

AI models have demonstrated unprecedented advantages in large-scale parameter tuning, nonlinear fitting, and real-time optimization [95]. In the context of the highly intricate architecture and nonlinear traits of the 6G networks [96], AI technology can gain a more profound understanding of the network's nonlinear features by learning from and dissecting extensive network data, and optimizing network parameters based on real-time data and scenario. Deep RL (DRL) has been deployed for the optimization of network parameters [97]. Specifically, each performance metric within the current communication network is considered a distinct RL state, and the adjustment operations carried out on network parameters are designated as diverse actions. In this procedure, by incessantly adapting the behavior of the RL model across different states, the model will receive feedback information from the environment to assess the merit of each action. RL models typically reinforce those actions that can yield higher reward returns.

2.2.4 *AI-based mobility management*

As early as 2015, 3GPP began researching how AI technology could be integrated into the 5G network. For example, the 3GPP SA2 and SA5 working groups (WG) launched projects on "enablers for network

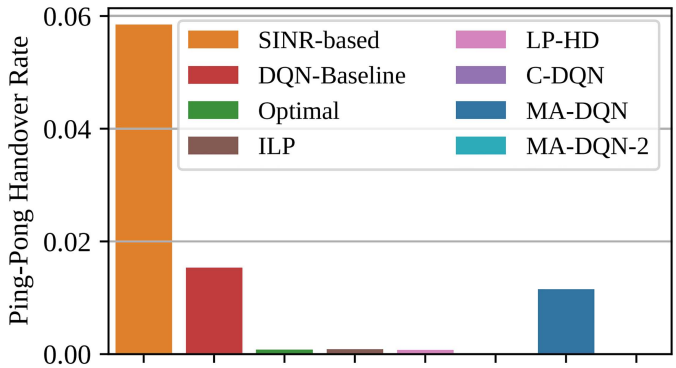

Figure 10 (Color online) Ping-pong handover rate performance comparison in [101].

automation" and "enhancement of management data analytics services", which made good progress [98]. Using technologies such as DL, AI is capable of predicting user behaviors, including moving speed and direction, thereby permitting the proactive selection and handoff between cells to diminish signal overhead. In ultra-dense network scenarios, frequent cell exchanges can lead to increased overhead and connection disruptions, adversely affecting QoS. The authors of [99] proposed the use of a GRU-based RNN to adaptively handle handoff decisions, significantly reducing handoff latency and improving network throughput. In [100], the authors suggested a method for predicting the next position of vehicles using AI in ultra-dense network scenarios, which allows for BS selection based on predicted positions, thereby reducing handoff frequency and failure rates, and saving handoff signaling overhead. This demonstrates that AI-based mobility management is feasible and practical.

On the other hand, through RL and other techniques, AI can dynamically optimize mobility management strategies based on network environment and user behavior, better addressing complex network environments. As wireless networks in higher frequencies require more dense deployment of BS, the authors of [101] addressed the issue of reduced network capacity due to frequent handoffs, proposing a centralized and multi-agent deep Q-network (DQN) algorithm. Both algorithms can find near-optimal user-to-BS allocations. As shown in Figure 10, the DQN algorithm significantly reduces or even completely avoids ping-pong handoffs compared to the baseline (conventional signal-to-interference-plus-noise ratio (SINR)-based handoff algorithm). The authors of [102] proposed an ML-based framework to optimize the handoff between sub-6 GHz and mmWave bands in vehicular networks, using RL to enable vehicles to predict and discover the optimal handoff target in real-time. The method reduces handoff latency, improves connectivity in vehicular networks, achieves seamless handoff even in high mobility scenarios, and enhances user experience.

2.2.5 *AI-based resource orchestration*

With hundreds of billions of devices being connected, the generated data are undergoing continuous explosive growth [103]. Merely strengthening the communication capabilities is insufficient to meet the real-time requirements for data processing in control application scenarios [104]. By harnessing the communication and computing capabilities at the network edge and on the device side, the entire process of data collection, processing, analysis, and decision-making can be accomplished closer to the device side, avoiding the latency drawbacks caused by congestion in the CN.

The joint scheduling of communication and computing resources founded on RL attains the collaborative and optimal allocation of communication and computing resources via the interaction and learning between the agent and the 6G network environment [105]. Confronted with the resource allocation difficulties within the space-air-ground integrated network in the 6G era, the authors of [106] put forward an RL-assisted bandwidth-aware virtual network resource allocation algorithm. This algorithm uses RL with a policy network for node embedding, preferentially handling high-bandwidth requests to meet strict user bandwidth demands. Given the relevant problems that exist within 6G mobile devices and communication systems, motivated by the concept of virtual network embedding, the authors of [107] pioneered a multi-objective aware dynamic resource scheduling algorithm designed for multi-layer computing networks, aiming to augment resource flexibility. They established a scheduling network underpinned by DRL and refined the learning process, thus proffering a sustainable resolution for resource scheduling

strategies. Regarding the circumstance that the applicability of the open radio access networks (O-RAN) in governing and optimizing the functions of the wireless access network within the 6G network has not been extensively explored, the authors of [108] developed low-complexity algorithms based on methods such as RL to address the problems of jointly optimizing traffic splitting and allocation, congestion control, and scheduling across different time scales. This effort provides insights for realizing a fully automated network with enhanced control and flexibility.

To address the challenges of computing resource demands and intelligent resource allocation in 6G network IoT services, the authors of [109] proposed an effective DRL-based solution for IoT resource expansion and service placement, and verified the effectiveness of its multi-application autonomous resource allocation via dataset simulation. In [110], the authors proposed a multi-task DRL method based on graph convolutional networks by introducing a joint network slicing and routing mechanism, achieving robustness in various network environments. In the study of [111], the authors established a model-free DRL framework by a hierarchical structure integrating modified deep deterministic policy gradient (DDPG) and double DQN to actualize an intelligent RAN slicing strategy with two-layer control granularity for maximizing QoS and slice spectrum efficiency. Moreover, a dynamic spectrum allocation scheme was proposed in [112], which utilized the advantages of the DQN and reduced the search state explosion by a reward and punishment framework to dynamically allocate unallocated resource blocks to mobile units and obtained the Pareto optimal solution of sub-problems via Chebyshev decomposition.

The AI-based resource orchestration still faces numerous challenges and opportunities in the future. For instance, the improvement of DRL algorithms is urgent to cope with complex and changeable environments. It is necessary to strive to enhance the adaptability to complex scenarios, meet the diverse resource requirements in various special scenarios of different industries, and improve performance evaluation and enhance algorithm interpretability.

2.2.6 *AI-based situational awareness and fault detection*

Network situational awareness technology endows mobile communication networks with real-time monitoring, prediction, and response capabilities. Its applications include network management and optimization, security monitoring, as well as fault detection and prediction [113]. AI-based situational awareness collects, stores, and analyzes vast amounts of network data via big data platforms and DL techniques, thereby providing precise threat intelligence and full traffic inspection [114]. This technology employs DL to precisely identify the states of various elements including network traffic, logs, and network key performance indicator (KPI), predict network development trends, and formulate accurate response strategies through the analysis of perceived information and predicted states [115], realizing the shift from "passive defense" to "active defense". The AI-driven threat representation and causal reasoning can identify and infer the origins of attacks, preventing the recurrence of similar security events.

The network fault analysis constitutes one of the crucial elements in the O&M of mobile communication networks. As the number of mobile devices increases and the network scale expands, the quantity of network alarms is growing explosively, and the relationships among them are highly intricate [116]. Employing AI to build a fault detection analysis model endows the model with the capacity for high-dimensional data analysis, enabling it to serve as a more intelligent and efficient tool for fault analysis [117]. By conducting a fusion analysis of these data, the system can effectively identify common issues within the network and perform accurate fault diagnoses. The results obtained after fault handling will be utilized as feedback to update and optimize the fault diagnosis model, allowing the system to operate efficiently under constantly changing network environments and business requirements [118]. The specific operational workflow is illustrated in Figure 11.

By integrating AI technology, we can establish a systematic and intelligent network fault analysis and tracing back scheme, which can provide automated, accurate, and flexible fault detection and tracing back capabilities and continuously enhance the efficiency of fault diagnosis and handling. Meanwhile, it can also adapt to the changes in network environment and business requirements.

3 Network for AI

A series of AI use cases and scenarios have been developed, and part of the standardization work has been accomplished under the leadership of 3GPP. This has led to the creation of various module-level "plug-in" AI functions in 5G networks. Although this approach is convenient, it still lacks considerations

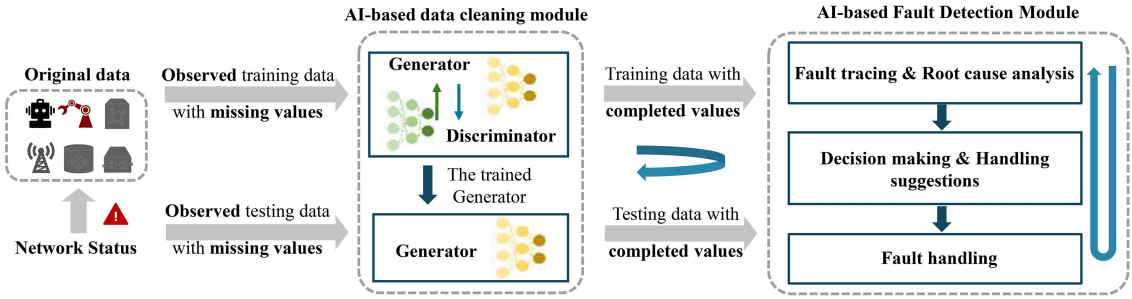

Figure 11 (Color online) AI-based fault detection.

of systematicity and interpretability, limiting the scalability of AI capability and the ability to model generalization.

Key technologies such as cloud-native architectures, software-defined information technology, and service virtualization have advanced significantly in 5G (e.g., CN supporting service-based architecture and cloud-based deployment of CN functions). 3GPP has also standardized the NWDAF as an AI-enabled component within the CN. This includes defining relevant mechanisms and processes while introducing standardized AI operational paradigms. However, these advancements remain constrained by several factors, including technological maturity, security concerns, and the complexity of system O&M. Consequently, the current access network architecture largely retains the traditional "siloed communication-technicalized BS design". In this architecture, BSs are highly tailored to communication services, lacking the flexibility to interface with emerging AI-driven services. From an end-to-end AI service perspective, there is no clear definition of AI functional components on the access network side. The integration and coordination of AI service functions between the CN and the access network remain unaddressed. This disconnect hinders the current network architecture from supporting the intelligent vision envisioned for 6G.

To support "native intelligence" and achieve ubiquitous general intelligence, the 6G network requires groundbreaking innovations in the network architecture to enable the endogenous embedding of AI capabilities within the network. This represents the core concept of Phase II (i.e., NET4AI) in the deep integration of 6G and AI, signifying a shift from a communication network solely providing connectivity services to a new network paradigm that offers a diverse array of services, including connectivity, computing power, data, and algorithms. This new paradigm would span CN, access networks, and terminals, to provide comprehensive support of AI.

3.1 NET4AI architecture over 6G

China Mobile [119] proposed a systematic 6G network architecture design that transcends connectivity in the industry, centered on the core concept of "three layers and six planes". The "three layers" refer to the physical resource layer, the network function layer, and the application&service layer. The "six planes" comprise the communication plane, the data plane, the computing plane, the intelligence plane, the security plane, and the management&orchestration plane. The first five planes are independent functional planes formed by the decoupling of the 6G network from end to end, providing corresponding capabilities and services. The management&orchestration plane realizes the flexible combination of functions across layers and planes, realizes the management and scheduling of multiple resources, and forms a complete service chain. This architecture focuses on multi-dimensional capability elements within the network. In [120], China Telecom proposed a ubiquitous and ultra-converge 6G network architecture, built upon the core concept of "three layers and three sectors". The "three layers" consist of the infrastructure layer, the network function layer, and the network operating system layer, while the "three sectors" embody data integration, intelligent simplicity, and trustworthiness. This architecture leverages the intelligent simplicity sector to integrate the intelligent brain, connections, and facilities across the three layers, thereby enabling intelligent inclusiveness. China Unicom [121] envisioned a future 6G network as intelligent, converged, green, and trustworthy. Its architecture is structured vertically into the land-based, air-based, and space-based communication layers and horizontally into three domains: sensing resource domain, function control domain, and service application domain. These layers and domains are interconnected by two chains: intelligent native and secure&trustworthy.

Huawei [122] unveiled a task-centric AI architecture of 6G networks, asserting that 6G networks will introduce new resource dimensions and support the coordination of multi-dimensional resources. This design allows 6G networks to natively support AI and achieve a transformation from session-centric to task-centric operations. CICT [123, 124] put forward the concept of multi-network integration and ubiquitous intelligence, envisioning that 6G will realize global three-dimensional deep coverage, develop body area networks centered on human beings, form multi-layer coverage with both breadth and depth and learn the ubiquitous intelligence system with multi-network integration. OPPO [125] suggested that 6G would redefine human interaction and AI, by making AI a ubiquitous infrastructure. The 6G system architecture will emphasize the coordination across intelligence, performance, and flexibility, and integrate deeply AI capabilities into the user plane, control plane, and function plane. More authoritative and official, The IMT-2030 (6G) Promotion Group in China envisioned the 6G network architecture in [34], as an open and innovative platform for information services, offering capabilities that transcend mere connectivity. These capabilities encompass computing power networks, trusted security, sensing, data services, and AI-based network intelligent autonomy.

Globally, Ericsson [126] has focused on cognitive networks by leveraging AI to achieve data-driven security and highly automated network operations. Qualcomm advocated that AI and ML would fundamentally transform the design and deployment of wireless communication and networking systems. As the leader of the EU's 6G project Hexa-X, Nokia [127] stated that AI would be the primary driving force behind the technological transformation of the NR interface for 6G. The company is creating unique AI use cases and scenarios and developing foundational AI technologies. LG Electronics is focusing on the research and development of 6G AI technology, anticipating that 6G systems will usher in the era of the "Internet of everything and environment" powered by AI. Japan's "Beyond 5G promotion strategy" project [128] aimed to introduce AI into network management, enabling network autonomy and driving network evolution towards intelligence.

Academic research has also proposed various innovative ideas for shaping the future 6G network architecture. The authors of [129] put forth a 6G network design philosophy of "Human-Machine-Thing-Spirit", emphasizing the inevitable trend of AI integration in 6G, leveraging cognitive enhancement and decision-making simulation to intelligently define network requirements, ensuring secure and reliable network transmission, and realizing intent-driven networks. In [130], a fusion of 6G RAN and AI was advocated, identifying four key characteristics: green, multi-dimensional, stereo, and full-scenario service. 6G would integrate communication, computation, control/caching, and AI to support full-application scenarios, and would be inherently secure [131]. The authors of [132] demonstrated a vision of integrating 6G and AI and studied implementation examples in wireless communications based on two classic AI algorithms: DL and RL. It was suggested in [37] that AI would arguably become the cornerstone of 6G with "intelligent inclusion" serving as an essential feature of 6G, enabling the network to provide intelligent services to all types of end users. The authors of [133] promoted that AI would be used in wireless communications and change the top-level architecture of wireless networks. The authors of [134] emphasized that the NET4AI paradigm should leverage the capabilities of 6G networks, such as precise wireless sensing, high data-rate transmission, and ubiquitous connectivity, for intelligence distillation at the network edge, hereby promoting edge AI and enabling ubiquitous and trustworthy AI applications in wireless networks. Among them, federated edge learning is expected to be an important learning architecture over 6G. 6G networks and AI will be highly integrated, and customized AI services will be provided on the wireless network side [135]. 6G would enhance its progress in intelligence, especially in edge intelligence [136]. Among them, integrating DT and edge networks is a promising attempt. The 6G Flagship research program was initiated in [137], advocating for realizing context-aware intelligent services and applications for human and non-human users through AI.

The industry has reached a core understanding of the intelligent design of 6G network architectures, emphasizing "distributed structure", "endogenous intelligence", and "integrated simplicity". Future networks will evolve into advanced, integrated platforms that surpass mere communication connectivity services, offering multi-dimensional services and corresponding network capabilities beyond communication. They will cater to diverse businesses by offering end-to-end lifecycle management services throughout their entire process. Additionally, the network's AI capability can be opened to the outside world, while external AI capability can be introduced into the network, enabling the crowdsourcing of AI capability.

We elaborate on the NET4AI architecture, i.e., the support provided by the other five capabilities/services in the network (i.e., communication connectivity, data, computation, security, and management & orchestration) for the AI capabilities/services, with an emphasis on their interaction with

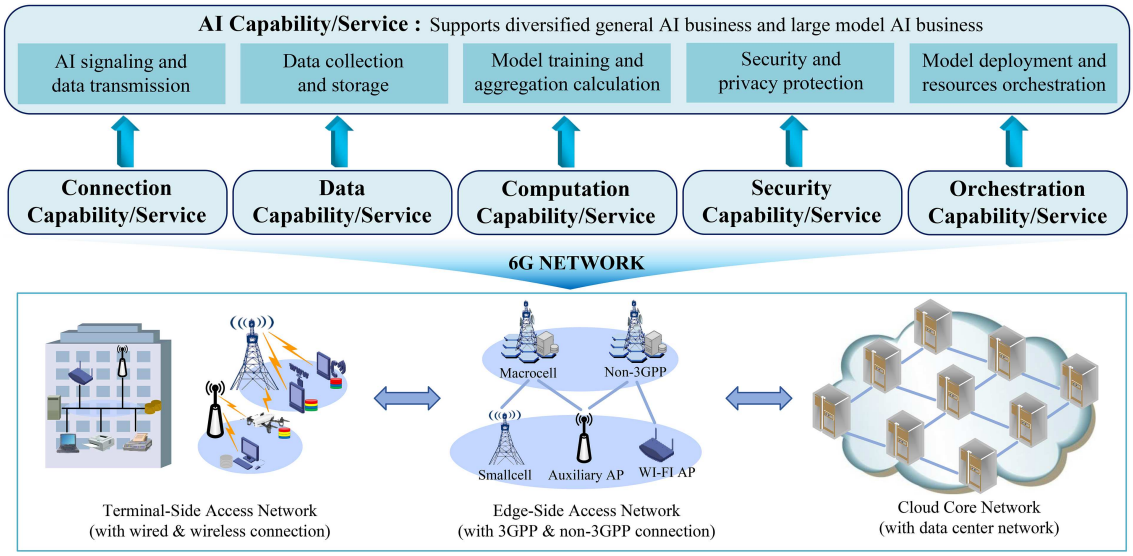

Figure 12 (Color online) Architecture diagram of NET4AI. The 6G network physically comprises three entity networks: the cloud CN (including centralized data/computation centers), the edge-side access network (including 3GPP and non-3GPP connection methods), and the terminal-side access network (including wired and wireless connection methods), as depicted in the lower half of the figure. NET4AI represents the mapping and supporting role of communication connectivity, data, computing, security, and management orchestration within AI, among the six capabilities and services (i.e., communication connectivity, data, computation, security, AI, and management & orchestration) abstracted from the 6G network's functional services, as indicated by the arrows in the upper half of the figure.

AI, as illustrated in Figure 12. NET4AI will provide architectural support for the implementation of AI-native 6G, achieving a deep integration of the network and AI.

3.1.1 *Connection for AI*

In 6G, the control plane and user plane will be further enhanced, aiming to deepen the service-oriented and open nature of network functions and provide support for emerging AI services. On the one hand, 6G needs to further enhance and optimize the service management and interface of the existing control plane functions to reduce the complexity of network functions and signaling interaction processes [138]. Additionally, 6G will extend the conventional control plane by integrating space-terrestrial ubiquitous access control management, enabling new functionalities and service-oriented offerings for session management and policy control across communication, computing, data, and other essential elements. On the other hand, the user plane will be enhanced by augmenting or introducing new, more generalized, and flexible user plane forwarding protocols that can perform agile service forwarding and processing based on communication, computing, and other elements. Leveraging microservice technologies and next-generation hardware support, e.g., high-performance programmable chips, cloud-native technologies, and the P4 (programming protocol-independent packet processors) advanced language, user plane functions (UPF) will be restructured into service-oriented modules [122,139]. This flexible invocation and efficient collaboration among functional modules will break the constraints of conventional layered protocols, enhancing data forwarding efficiency and bandwidth utilization.

Considering the control plane and user plane separation principle [130], the types of AI service-related data carried by communication connections can be naturally categorized into two major groups: AI signaling and AI data. Connection for AI embodies providing connectivity support for AI signaling and data.

• **AI signaling.** AI signaling is used for transmitting AI-related control messages, e.g., AI service request/response messages; request/response messages related to computing power required for AI analysis, etc. It can also include auxiliary information for collaborative AI analysis among multiple entities.

• **AI data.** AI data are used to transfer AI input data, generally referred to as input vectors for invoking an AI model or represented as inputs in an AI algorithm. It is also used for transmitting AI outputs, generally called the output of the invoked AI model/algorithm. It can be used for transmitting AI models, e.g., transmitting a neural network model usually means transmitting its parameter weights (which can be represented as vectors) and its network structure (which can be represented as scalars). It

can be further used for transmitting AI training data. The data can often be usually huge and dedicated to the training phase of intelligent models.

For AI business scenarios and related connectivity performance requirements, 3GPP TS 22.261 [140] (based on the conclusions of TR 22.874 [141]) also has some KPIs requirement level analysis and results for 5G systems, mainly on the impact of conventional connectivity performance metrics, such as throughput, latency, and reliability; in other words, the current 3GPP standard only analyzes the impact of the 5G system (mainly the CN portion) in terms of how it supports different AI services/applications from the functional and performance perspectives. From an overall end-to-end process analysis perspective, AI signaling and AI data involved in AI services/applications necessitate transmission via the air interface. However, in the current standards and architectures, no relevant modules guide the transmission process or standardize the transmission content.

3.1.2 *Data for AI*

6G networks necessitate a novel data capability that differs from the conventional user plane. This capability must adhere to data regulation and supervision requirements, while enhancing data analysis and processing efficiency, addressing data management challenges, and supporting AI services. This data capability supports the realization of information collection, cleaning, and storage, and enables the complete lifecycle management of data service processes while ensuring data privacy and security. By leveraging a hybrid data processing and storage model that combines centralized and distributed approaches, the 6G network can flexibly address diverse scenario requirements: (1) for sensitive or high-privacy-level data, local processing and storage can be performed at the RAN/CN and terminal sides to ensure no data leakage; (2) global and comprehensively analyzed data, on the other hand, is processed and stored centrally in the data center. Through tight cooperation between the central data center and edge nodes, 6G can efficiently manage massive amounts of data originating from various technological and business domains while safeguarding privacy.

Data capability provides a solid data foundation for 6G's endogenous intelligent AI services [142]. Several publicly available datasets have been widely utilized, such as DeepMIMO [143] and RadioML [144]. The DeepMIMO dataset, based on ray-tracing simulations of real wireless communication environments, provides MIMO channel data and is extensively applied in scenarios like channel modeling, beamforming, and user location in 6G networks. By contrast, the RadioML dataset encompasses radio signal data with multiple modulation schemes and is mainly used for signal modulation recognition, interference detection, and spectrum management. Moreover, the "AI/ML in 5G challenge" dataset [145] released by ITU covers a variety of communication scenarios in 5G networks, offering crucial support for research on network optimization, traffic prediction, and resource allocation. These datasets not only lay a solid foundation for AI research in the communications field but also provide unified benchmarks for validating and comparing algorithm performances.

Data for AI aims to efficiently support end-to-end data collection, transmission, storage, and sharing, addressing how to facilitate, expedite, and securely provide data to AI functions within or external to the 6G network. Depending on the potential functional scope, the data support provided by data for AI should encompass five components. (1) Data collection/distribution. This component provides a basic publish-subscribe mechanism for data producers and consumers, enhancing data collection/distribution efficiency and supplying samples for large-scale training required by AI models to achieve optimal results. (2) Data security and privacy. This component leverages security and privacy protection technologies to provide high-quality, trusted data services tailored to user and network needs, ensuring user and network privacy protection and data security, immutability, and traceability. (3) Data analysis. This component utilizes AI models, algorithms, and knowledge to provide statistical information, predictive insights, network anomaly analysis, and optimization suggestions, enhancing the data consumption experience for internal and external network functions and related intelligent applications. (4) Data preprocessing. This component applies generic tools for format conversion, noise reduction, feature extraction, and other preprocessing tasks to collected data, fulfilling AI model training requirements for input data and facilitating the management of multi-dimensional complex data. (5) Data storage. This component stores and retrieves collected data, providing storage support for data services related to data security and privacy, data analysis, or data preprocessing, and further establishing a data foundation to support AI services.

Through standardized data service functions, a comprehensive data service process can be established

within the 6G network. As researchers often need to construct dedicated datasets for specific application scenarios, data for AI can provide valuable guidance. Take the construction of communication datasets as an example. Firstly, data sources can be selected from collected data in real-world communication networks or generated simulation data using high-fidelity simulation tools (such as NS-3 or OMNeT++). Real-world data can reflect actual scenarios but may involve privacy and security issues; whereas simulation data offer higher flexibility and can be customized to specific scenarios according to particular requirements. Secondly, data preprocessing and labeling are two essential steps. Communication data usually contain a large amount of noise, thus requiring preprocessing operations, such as cleaning and normalization, and accurate labeling based on research objectives (e.g., channel state prediction or interference detection). When using real-world user data, strict compliance with data privacy protection regulations is mandatory. Techniques such as anonymization or differential privacy should be adopted. Last but not least, researchers are encouraged to make datasets publicly available under the premise of protected privacy, to promote collaborative research and algorithm comparison within the field. It is recommended to use a unified storage format and standard to facilitate data sharing and reproducibility of research results.

3.1.3 *Computation for AI*

The evolution of 6G will bring significant advancements in computation, including computational sensing, control, and execution. Computational sensing will gather information about the ubiquitous computing resources distributed across the spectrum, from central clouds to terminal devices. Computational control will cater to computing service demands by providing intelligent strategies for pervasive computing power control and resource scheduling. Computational execution, guided by these strategies, will deliver services such as session management, traffic control, and policy execution.

Computational capabilities must synergize with communication connectivity to optimize the efficiency and energy consumption of emerging services like AI. This integration necessitates the development of a unified environment that ensures continuous operation of computing services across terminals, edge nodes, networks, and clouds. Such an environment will enable the dynamic selection of optimal computing nodes for computational tasks, considering factors like latency, bandwidth, computing power, and energy efficiency. This unified framework will allow computational tasks to transition among terminals, networks, edge nodes, and clouds [146]. This computational power, along with the AI algorithms or functionalities, serves the network or devices to improve network operations but is also potentially exploited through unified interfaces to serve upper-layer applications. Computational power can be categorized into three types: dedicated computing power for network elements, distributed external computing power, and endogenous computing power within the network.

(1) Network element dedicated computing power for AI. This type of computing power typically consists of computing and storage units composed of specialized processors and programmable devices. It serves network elements in mobile communication networks (BS or CN). It is primarily used for customized AI application services (i.e., AI4NET) to enhance communication performance or optimize network operations, like CSI feedback, channel estimation, and beamforming. Due to resource restrictions, it cannot support AI services and applications requiring large-scale computations and training. It is challenging to support third-party applications.

(2) Distributed external computing power for AI. This type of computing power typically exists in the form of distributed edge computing/MEC, primarily utilizing general-purpose central processing units (CPU), high-performance graphics processing units (GPU), and programmable acceleration cards. External computing power shifts computations from centralized data centers to the edge of the access network, enabling network optimization and supporting industry applications with high computational demands and stringent latency requirements, such as video acceleration and augmented reality (AR)/VR scenarios. Due to its external nature, it requires management through unified network functions, introducing some latency that may impact extremely latency-sensitive AI services and applications. Moreover, this external overlay approach may not optimize resource utilization, hindering efficient construction, deployment, and usage of AI services.

(3) Distributed network-native computing power for AI. In the vision of future networks, each network element will possess control, forwarding, and computing capabilities, with numerous computing nodes deployed throughout the network. This computational power, known as endogenous computing power, promotes the development and deployment of endogenous AI, enabling large-scale intelligent

distributed collaborative services. It compensates for the shortcomings of external computing power, promptly responding to changes in mobility and networks while fostering the emergence and development of future AI applications, such as immersive cloud XR, holographic communication, sensory interconnection, communication-sensing integration, and DT.

3.1.4 *Security for AI*

Some security vulnerabilities have remained after the design of 2G–5G systems, and conventional security measures have often failed to ensure safety. 6G networks necessitate establishing an inherently trustworthy and self-driving security system at the core of the system. Rapid advancements in cloud computing, big data, and AI technologies offer technical support to build a 6G network security system. In recent years, trusted endogenous security has emerged as a new security approach for 6G, characterized by four main features: collaboration, intelligent proactive defense, trustworthiness, and privacy protection. The concept of trusted endogenous security also applies to ensuring the security of AI.

(1) Ubiquitous collaboration. Collaboration has become a technological approach to enhancing communication capabilities. Intelligent ubiquitous collaboration will enable NET4AI to adapt to various complex and dynamic environments and scenarios while improving the robustness of the AI plane to identify potential malicious attacks or abnormal behaviors accurately. Future security defenses will shift from isolated to highly efficient collaboration, achieving individual collaborative defenses between entities or intelligent agents, inter-layer defenses between protocol layers or architecture layers, and inter-domain defenses between network space domains.

(2) Intelligent active defense. The rapid development of 6G networks and AI technology presents opportunities for intelligent security defenses. The intelligent, proactive defense will enable NET4AI to continuously learn and adapt while automatically adjusting defense strategies based on changes in the network environment and AI security threats to counteract attacks promptly. This transition represents a shift from passive protection to proactive sensing and defense [147].

(3) Trustworthiness. 6G will further evolve towards trustworthiness capabilities, achieving a comprehensive endogenous security system encompassing security (i.e., confidentiality, integrity, availability), reliability, resilience, and safety [148]. This can provide a trusted ecosystem for NET4AI. Furthermore, flexible and reliable access control [149] and effective identity authentication [150] can, to some extent, prevent unauthorized access from impacting the endogenous trusted environment of NET4AI.

(4) Privacy protection. The deep integration of AI, big data, and 6G networks will exacerbate the challenges of data privacy protection. NET4AI must handle massive business and personal data while preventing privacy leakage. Emerging distributed ML, while achieving distributed ubiquitous intelligence and partially avoiding privacy leakage caused by AI model training, still has security risks [151]. Therefore, NET4AI will integrate distributed ML with existing privacy protection methods at multiple points to enhance data privacy protection and build an efficient and secure data ecosystem.

3.1.5 *Orchestration for AI*

6G is considered to possess an extremely high level of network autonomy, which means that the management & orchestration capability of the network can conduct unified and dynamic orchestration and scheduling for various functional components of service requirements, thus enabling the self-optimization, self-evolution, and self-healing of the network. The 6G system will introduce external applications and service demands while exposing internal diversified resources and functions. This requires flexible cross-functional flow and scheduling of multi-dimensional resources, such as computing, communication, and storage within the network, offering customized and personalized application services based on virtualization and microservice technologies. For instance, AI capabilities and analyzed data within the network can be opened to third parties to provide services and necessary support, a crowdsourcing behavior that necessitates the support of management and orchestration functions. Furthermore, through joint analysis and prediction of computing resources and network traffic, containerized gNBs can automatically scale up or down based on predicted network traffic loads, enabling dynamic resource allocation and network energy savings. Therefore, management and orchestration capabilities are closely related to all other capabilities, orchestrating and managing resources and functions efficiently and flexibly across all functional layers of the 6G network, achieving joint optimization.

Orchestration for AI signifies efficient support for deploying, operating, and optimizing AI services within the network through orchestration. It automatically configures and orchestrates network resources

based on the specific requirements of AI services, ensuring that AI applications obtain the necessary computational power, storage, network bandwidth, and other resources. Additionally, this orchestration capability enables real-time monitoring of AI service performance, dynamically adjusting resource allocation to accommodate fluctuations in service loads, thereby guaranteeing the stability and efficiency of AI services [152]. This seamless integration and support facilitate smoother AI service operation within the network, providing robust network guarantees for various intelligent applications. Specifically, this manifests in the following three aspects.

(1) Automated deployment and resource allocation. The orchestration capability swiftly responds to AI services' network resource demands through automation. Upon AI service deployment, the orchestrator intelligently analyzes the current network resource utilization, automatically configuring network devices, establishing links, and deploying services. This automated process enhances deployment efficiency, mitigates human errors, and ensures rapid AI service launch and stable operation.

(2) Dynamic optimization and elastic scaling. As AI services operate, their resource requirements may fluctuate. The orchestration capability features dynamic optimization and continuous monitoring of AI service performance, including CPU usage, memory consumption, network bandwidth, and other key metrics. Upon detecting resource shortages or surpluses, the orchestration system automatically adjusts resource allocation, such as adding compute nodes, expanding storage capacity, or optimizing network bandwidth, to meet AI services' dynamic demands. Furthermore, the orchestration system supports elastic scaling, automatically resizing resources based on AI service loads, ensuring service stability and efficiency.

(3) Fault self-healing and intelligent O&M. By integrating monitoring, alerting, and troubleshooting functions, the orchestrator promptly identifies and locates faults within the network, automatically triggering recovery mechanisms. This fault self-healing capability minimizes service disruptions for AI services, enhancing service availability and reliability. Additionally, the orchestration facilitates intelligent network O&M, empowering network administrators to manage network resources and AI services more efficiently, reducing O&M costs, and improving efficiency.

3.2 Key enabling technologies

This subsection describes the key enabling technologies for NET4AI. Each of the five capability/service planes in the future 6G network architecture, as presented in Subsection 3.1 shall contain its own enabling technologies, as categorized in Figure 13. In the following, we focus on those closely related to the needs of AI, introducing their research background, significance, and progress.

3.2.1 *Distributed intelligence and FL*

The NET4AI architecture must address the dual demands of low latency and high reliability for communication and computing, as well as growing concerns about privacy. Distributed intelligence and FL are the key technologies that can solve these challenges. Conventional centralized architectures in wireless networks can no longer meet low-latency, high-reliability communication and computation demands. Additionally, conventional centralized architectures are inefficient in supporting the ubiquitous intelligence required by future wireless networks. This trend towards deploying intelligent decentralized elements has become apparent, with intelligence gradually shifting from centralized to distributed systems. Therefore, introducing distributed intelligent computing architectures is necessary to fully utilize the multi-dimensional data and computing resources held by user terminals and nodes, enabling intelligent connectivity.

Wireless distributed intelligence refers to organizing AI tasks in a distributed manner within wireless networks and incorporating collaborative AI and ML without uploading all raw data to a central cloud. This approach alleviates the network bandwidth pressure and reduces the central cloud's computational burden. Moreover, as the computing power of smartphones and IoT devices increases and users become more concerned about data privacy, mobile devices can handle more AI computations and training. This proximity of AI to users and data allows for faster iteration and upgrade cycles, efficiently supporting the pervasive intrinsic intelligence of 6G networks.

Distributed intelligence is a cornerstone technology for realizing the intrinsic intelligence envisioned for 6G networks. By adopting a decentralized architecture, it disperses network functions and resources across multiple nodes or devices, thereby eliminating reliance on a single central node to manage the entire network. Each node in a distributed network operates with a degree of autonomy, enabling independent

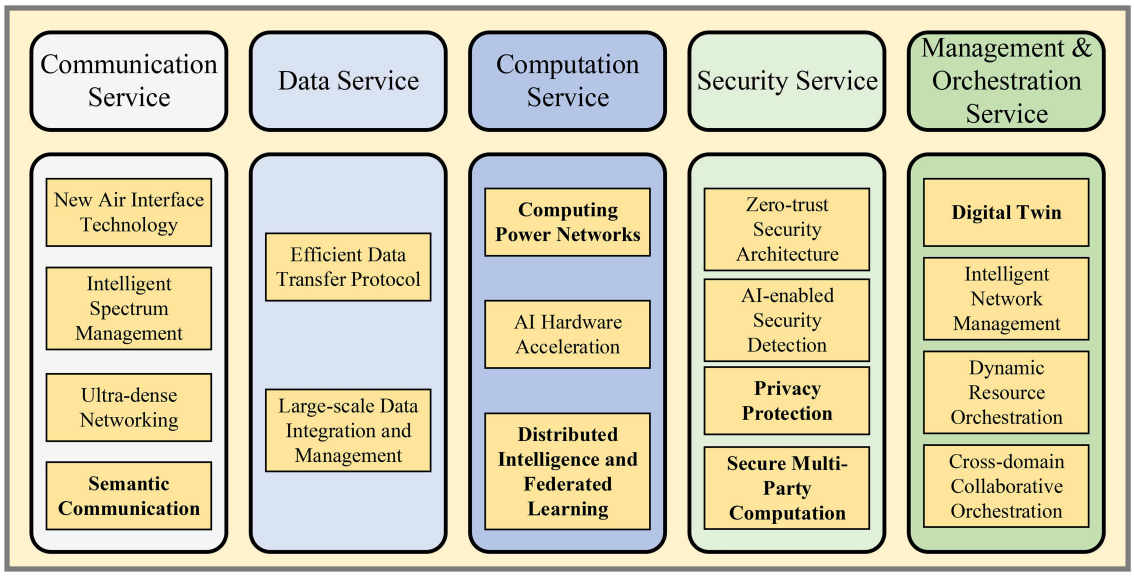

Figure 13 (Color online) NET4AI enabling technology.

decision-making and task execution without the need for constant communication or coordination with other nodes. This autonomy enhances network flexibility, making it more adaptable to dynamic environmental conditions and evolving demand patterns.

In distributed networks, data are stored across multiple nodes, and computational tasks are processed collaboratively among them. This decentralized approach to data storage and computation improves network scalability, accelerates the training and inference of LAMs, and enhances overall system efficiency. Moreover, distributed intelligence facilitates online learning and continuous optimization by enabling real-time data collection, processing, and model updates, which improves the performance and accuracy of AI systems. Distributed AI technology addresses the limitations of centralized AI systems, such as communication bottlenecks and data privacy concerns. It reduces communication overhead and ensures data security by keeping sensitive information closer to its source. Distributed AI is poised to create a new intelligent ecosystem for 6G, fostering a more scalable, efficient, and privacy-preserving network paradigm. As shown in Figure 14, the wireless distributed intelligence architectures include as following.

(1) FL. FL is a typical paradigm of distributed intelligence, which enhances privacy by training AI models on local devices and only sharing model parameters instead of raw data [153]. During training, each client node trains the model locally and uploads the model weights regularly. A central node aggregates these weights and feeds the aggregated weights back to the clients for the next round of training or local inference. A typical example of FL in wireless networks is illustrated in Figure 14(a). The process begins with downloading the global model ω^t for the current round t to local devices. Each device k trains the model on its private dataset D_k, generating a local gradient update $\boldsymbol{g}_k^{(t)}$, which is then uploaded to the parameter server. The server applies the federated averaging (FedAvg) algorithm to aggregate these updates into a global gradient $G^{(t)}$, subsequently updating the global model to ω^{t+1} for the next round. This process continues until model convergence. FL can be divided into horizontal FL and vertical FL. Horizontal FL emphasizes inter-source cooperation and data sharing, whereas vertical FL facilitates collaboration and information exchange across distinct levels within the same data source.

• **Horizontal FL.** Also known as sample-based FL, horizontal can be applied in scenarios where the datasets of the various participants in FL have the same feature space but different sample spaces. Through horizontal FL, different participants can jointly train a model based on local data sets while protecting privacy and improving the accuracy and generalization of the model. Horizontal FL enables real-time learning on distributed edge devices, allowing for timely model parameter updates to constantly changing environments and data. It supports collaboration across devices with varying computational capabilities and data distributions, catering to the diverse or personalized needs of 6G network terminals [154]. It also supports over-the-air computations to achieve excellent scalability [155, 156], tolerates imperfect wireless transmissions [157], and can be potentially implemented in a fully decentralized fashion [158].

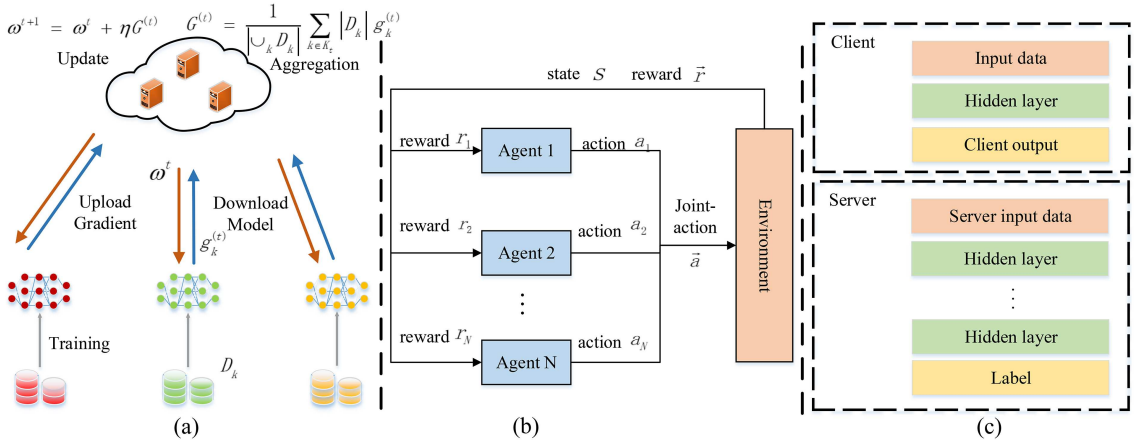

Figure 14 (Color online) Illustration of examples of (a) FL, (b) MARL, and (c) SL in distributed intelligence.

The authors of [159] considered limited resources in wireless networks, and analyzed the convergence of FedAvg in horizontal FL under full and partial device participation, providing important theoretical support for the application of horizontal FL in wireless networks. In [160, 161], the authors represented the partial device participation as a probability distribution and eliminated the device selection bias by modifying the aggregation rule in FedAvg to ensure the unbiasedness of gradient updates. This allows for a more stringent convergence bound and is committed to the efficient implementation of horizontal FL. The authors of [162, 163] provided a variety of solutions to guide the efficient application of FL in wireless networks, such as layering, aggregation frequency optimization, device selection, and communication compression, in support of low-carbon green economics. In [164], the authors focused on the shortcomings of horizontal FL in dealing with the heterogeneity of device data, introduced a personalized federated learning (PFL) method, and provided a theoretical basis for the application of FL in multiple fields to realize personalized models. The authors of [165] categorized PFL techniques based on key challenges and strategies, highlighting key ideas and future research directions in PFL architecture, realistic benchmarks, and trustworthy methods. In [166], the authors considered the application of horizontal FL in the medical field to solve the problem of collaborative research and patient health testing among multiple hospitals, and proposed a blockchain-based orchestration framework to enhance medical data privacy. The authors of [167] focused on IoT healthcare and proposed an intelligent medical system agent architecture that integrated horizontal FL and blockchain technology to achieve efficient and secure processing of medical data. The authors of [168] proposed a new FL framework for cross-domain prediction for intelligent manufacturing, designed to quickly adapt to new products, processes, and applications with limited training data. Horizontal FL has important application prospects in providing efficient and personalized AI services in wireless networks, coordinating research with multiple hospitals in healthcare, and realizing cross-application in intelligent manufacturing. However, it also has defects such as slow convergence speed and susceptibility to data heterogeneity and poisoning attacks [169, 170].

• **Vertical FL.** Vertical FL refers to a setup where datasets have different features but the same samples/users. It is suitable for scenarios where there is a high overlap among participants but low or no overlap of their features. During the training process, participant data must first be aligned through sampling to ensure overlapping datasets across users. Subsequently, the model is trained locally, and the resulting parameters are uploaded to a central server [171]. Vertical FL leverages correlations between diverse data sources to achieve data augmentation, enabling complementary and enriched datasets that enhance the generalization and robustness of AI models. By sharing model parameters, vertical FL facilitates cross-level learning and knowledge transfer, allowing data sources to exchange learning experiences and improve overall model effectiveness. To tackle the problem of user diversity, a novel approach for a vertical collaborative recommendation system based on clustering algorithms and latent factor models is put forward [172]. This method effectively enhances the accuracy of the recommendation results. In [173], the authors developed the idea of intent-hidden vertical FL for multi-party collaborative training of disease prediction models in medical diagnosis, and built a security protocol using homomorphic encryption for data privacy protection. The authors of [174] proposed a vertical federated DL based on the spatial-temporal long- and short-term networks for intelligent city traffic in MEC environments. Lever-

aging diverse sample features (e.g., different regions) within the same sample space (e.g., a specific traffic scene and time range) enables the construction of a federated model for traffic flow prediction. Vertical FL can uncover relationships between different features of data samples, making it highly valuable for applications such as recommendation systems and traffic prediction. Research on vertical FL remains in its infancy, with limited theoretical advancements to date.

(2) Multi-agent reinforcement learning (MARL). In a shared environment, multiple learning agents pursue their own rewards. 6G networks, supporting diverse AI applications like smart cities and intelligent transportation, can benefit from MARL. With its adaptive decision-making, MARL enables agents to autonomously optimize and adjust, enhancing user experience across various applications. The authors of [175] introduced a decentralized MARL algorithm based on advantage actor-critic to address large-scale traffic signal control, effectively tackling congestion in intelligent transportation systems. In [176], a multi-agent recurrent DDPG algorithm was proposed to manage traffic light phases in real-time using Internet of Vehicles data, mitigating congestion in complex and partially observable road environments. In [177], a partially observable Markov game was used to model connected autonomous driving, resulting in a multi-agent learning platform for interconnected autonomous vehicles. The authors of [178] focused on collaborative computing in low earth orbit (LEO) satellites, developing a real-time offloading decision model optimized through MARL with a penalty function. Additionally, an MARL-based online production scheduling method was proposed in [179] for smart factories, while multi-agent DRL was applied in [180] for multi-channel access and task offloading in MEC-enabled Industry 4.0.

(3) Split learning (SL). Due to the limitations of resource-constrained devices that cannot support complex DL models or FL collaboration [181], SL has emerged as a viable solution [182]. SL divides ML models, distributing different parts between the client and server, ensuring that original data remains secure [183]. While high-computing tasks are handled by powerful central nodes, lighter data processing layers are retained on the terminal where the data are stored. As an alternative or complement to FL, SL reduces the processing burden on resource-limited devices and holds the potential for advancing distributed intelligence in 6G. The authors of [184, 185] introduced SL in healthcare applications, enabling multiple entities to collaboratively train DL models without sharing sensitive raw data, achieving promising performance results on medical datasets. In [186], SL was explored for multi-institutional collaborative training, demonstrating improved client efficiency and more reliable privacy protection. To address issues such as overfitting and slow convergence in the original sequential training method, the authors of [187] proposed a parallel SL method. The method mitigated overfitting through adaptive training batch adjustments and node layer synchronization, improving training efficiency while maintaining privacy. In [188], SL was applied to wireless networks to leverage terminal privacy data in IoT. A cluster-based parallel training method was proposed, optimizing joint cutting layer selection, device clustering, and spectrum allocation to reduce training delays effectively. However, SL remains vulnerable to data heterogeneity, complex training methods, overfitting, slow convergence, and potential failure to converge.

(4) Summary and lessons learned. The above-mentioned technologies can all contribute to distributed intelligence. FL is suitable for AI services in scenarios with data privacy requirements, and applicable to various emerging intelligent demand applications of 6G owing to its compatibility with other intelligent learning paradigms. MARL is in line with the needs of online learning in complex environments and completes AI tasks in an observable and interactive environment. SL is suitable for collaborative learning of DL models in resource-constrained devices, and can adapt to different levels of resource constraints through multi-level model splitting. Distributed intelligence may expand into additional research directions. These include: (1) the adoption of error correction and retransmission mechanisms, such as advanced error correction codes and adaptive retransmission strategies, to ensure reliability in unreliable communication environments; (2) the implementation of secure protocols and privacy protection measures, including data encryption, identity authentication, and access control, to safeguard user data from unauthorized access and tampering; (3) the efficiency of data/model parameter exchange to enhance the energy efficiency of learning algorithms; and (4) fairness through incentive mechanisms for participating devices [189].

3.2.2 *DTs for 6G network native AI*

Conventional communication systems relying on historical data and static models struggle to adapt swiftly to dynamic network changes, making real-time adjustments difficult. These limitations hinder

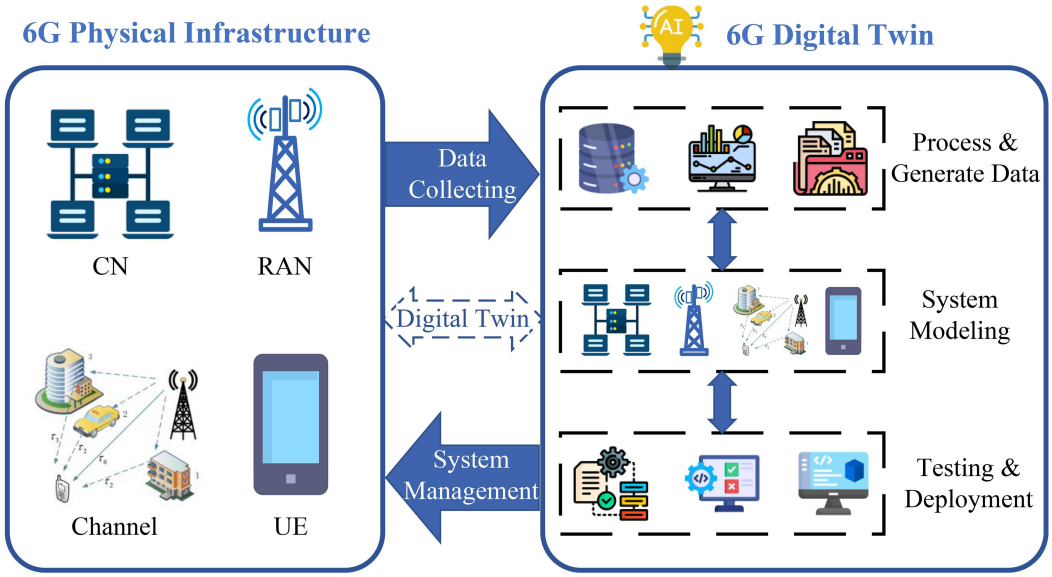

Figure 15 (Color online) DT for AI.

the adaptability, reliability, and flexibility of the systems. DT technology has been introduced into the communication domain to address these challenges, offering a flexible, efficient, and precise approach to resource optimization. Adopting DT not only enhances system agility but also introduces innovative models for network operation, significantly reducing research and development costs and risks [190]. Figure 15 depicts the relationship between 6G physical infrastructure and DT for AI.

The concept of the DT was first proposed in [191]. By creating a virtual replica of a physical object, DTs establish a bidirectional connection between the physical and digital worlds, enabling real-time reflection of the physical entity's operational state and the simulation of its future behavior [192]. In 6G communication systems, DT technology facilitates the construction of comprehensive network digital replicas to monitor network devices in real time, predict network states, and optimize resource allocation [193].

The implementation of DT relies on the synergy of several key technologies, including high-precision data collection, real-time modeling and simulation, bidirectional data interaction, and intelligent analysis and optimization. In communication systems, distributed sensors continuously gather diverse data about network device statuses, traffic dynamics, environmental parameters, and user behavior patterns. These data are transmitted to the DT platform via high-speed, low-latency communication links. The DT platform employs physical models and data-driven ML algorithms to construct high-fidelity virtual replicas that comprehensively simulate physical entities' operational state and dynamic changes [194]. With bidirectional data interaction, physical systems and virtual replicas achieve real-time data synchronization. When a network traffic surge is anticipated, the DT can proactively adjust load distribution strategies, optimize spectrum and power resources, and alleviate stress on network devices. The integration of edge computing further enhances the efficiency of this technology by distributing modeling, simulation, and optimization tasks to edge nodes closer to physical entities. DT enables intelligent management and dynamic optimization of complex networks in 6G communication systems, providing robust support for efficient resource utilization, network stability, and intelligent upgrades [195].

The 6G network aims to achieve ultra-high capacity, ultra-short-range communication, beyond-best-effort service, high-precision communication, and the integration of multiple communication types [196]. However, this vision introduces security, spectrum efficiency, intelligence, energy efficiency, and affordability challenges. The emergence of DT technology offers promising opportunities to address these challenges. DT provides a virtualized counterpart for 6G networks, enabling comprehensive network traffic monitoring and analysis. Leveraging feedback from the virtualized network, 6G systems can enhance their security by preparing for potential threats in advance. Additionally, DT enables the automation of demand identification and service provisioning by analyzing communication data to discern usage patterns and rules. Predictive insights into communication demands allow 6G networks to reserve resources, such as spectrum, in anticipation of future needs. Moreover, the integration of DT technology empowers 6G to support innovative services, including AR/VR, and autonomous driving. By addressing key chal-

lenges in security, spectrum efficiency, intelligence, energy efficiency, and customization, DT technology redefines and accelerates the development of 6G networks [197].

Significant research has integrated DT technology with 6G networks to enhance network performance. For instance, DT has been proposed for managing metasurface reflectors in 6G terahertz communications [198]. By modeling, predicting, and controlling the propagation characteristics of indoor signals, DT maximizes the system's terahertz signal-to-noise ratio (SNR). Additionally, Cui et al. [199] introduced the concept of DT wireless networks by integrating DT into wireless networks. This approach shifts real-time data processing and computation to the edge plane and leverages DT to mitigate the unreliability of long-distance communications between end users and edge servers in 6G networks.

One of the critical capabilities of 6G networks will be supporting massive-scale terminal devices that generate extensive data traffic. For example, the authors of [136] designed a DT model for edge computing in 6G wireless networks. By incorporating mobile users' dynamic network state and DT, the study addressed edge association problems using DRL and transfer learning, successfully reducing average system latency and improving resource utilization. Similarly, the authors of [200] proposed a DT-assisted task offloading method, creating DTs for all device states to achieve lower latency and power consumption in the network. Furthermore, the authors of [201] investigated dynamic resource allocation mechanisms for DT-based services in the 6G IoT. The study improved resource allocation efficiency by establishing service function chains with DTs and employing a collective RL method.

In addition, the DT channel (DTC) has been proposed as a digital virtual mapping of a wireless channel that reflects the entire process of channel fading states and variations in the physical world [202]. DTC can be applied in various typical 6G usage scenarios, offering significant performance enhancements in data rate, latency, spectral efficiency, energy efficiency, and link reliability [203]. A novel DTC implementation platform was proposed, including data acquisition, environment sensing, and reconstruction module to build the relationship between environment and channel for channel prediction, communication decision, and interaction [204, 205]. A cluster-nuclei based channel model was proposed to analyze the mapping relationship between clusters and scatterers in the propagation environment, enabling channel reconstruction across various scenarios [206]. In addition, a radio environment knowledge (REK) pool was proposed to serve as a specific enabler for DTC. An electromagnetic wave property-inspired REK construction method was proposed, bridging the gap in the interpretable mathematical representation of the relationship between environmental information and the channel [207].

The ITU documentation defines the requirements and architecture of DT networks DTNs [208], emphasizing their core role as virtual representations of physical networks. DTNs enable physical network analysis, diagnosis, simulation, and control. The ITU proposal introduces a "three-layer, three-domain, double closed-loop" architecture design, highlighting applications in complex network operations, optimization, innovation, and security strategy drills. For instance, DT optimizes computation offloading in industrial IoT in uRLLC links. Integrating DT with blockchain and FL technologies in wireless communication strengthens system security and privacy protection [209]. In industrial IoT environments, the authors of [210] minimized computation offloading delays in uRLLC links by leveraging DT-enabled wireless edge networks. To enable wireless communication systems based on proactive online learning, a DT design framework was introduced in [211]. This framework considers aspects such as twin object access, security and privacy, and air interface design. Additionally, in the work of [212], a blockchain-empowered FL framework was proposed to enhance communication security and data privacy within DT-enabled edge networks. The study incorporates asynchronous aggregation mechanisms and DT-driven RL to optimize user scheduling and spectrum resource allocation.

An innovative vehicular edge computing network combining DT and multi-agent learning was proposed in [213], which uncovers potential matches for edge services among large-scale vehicles, reducing the complexity of service management. The authors of [214] presented a DT-assisted real-time traffic data prediction method capable of delivering accurate short-term forecasts for traffic flow and speed. A DT-based load-balancing prediction model for autonomous vehicles was introduced in [215], which accurately predicts road network conditions while ensuring high data transmission security. In [216], the authors explored the use of DT in vehicular edge computing, developing models that reflect the real-time state of the vehicular environment. The study introduced two metrics, quantum DT and cognitive DT, to evaluate the quality and cost of DT models for edge nodes.

3.2.3 *Computing power network and AI*

The NET4AI architecture requires powerful computing service capabilities because wireless networks show great interest in computing tasks related to high-performance computing, such as ultra-large-scale data processing and DL. The computing power network is a key enabling technology for computing services [217]. It achieves ubiquitous computing interconnection through mutual sense and collaboration between the network and computing resources [218], harmonizes cloud, edge, and terminal resources, and serves all computing tasks in the network, especially AI-related computing services. The computing power network is a new type of information infrastructure. As shown in Figure 16, it supports computing power discovery, computing power management, task offloading, and network computing [218,219]. It connects geographically distributed computing power center nodes through new network technologies [217], and computing power, data, and application resources are gathered and shared [220].

The technical principles of the computing power network have three aspects: (1) real-time and accurate computing power discovery; (2) flexible and dynamic service scheduling; and (3) consistency of user experience. Currently, the computing power network has gained wide recognition in both academia and industry and is classified into the following three categories.

(1) Computational power measurement and modeling. This is a foundational technology for providing computational power services [218, 221]. In the future, computational power providers in computational power networks will not be confined to a specific data center or computing cluster. They will include ubiquitous computational power from cloud, edge, and end devices. Efficient sharing of this ubiquitous computational power through network connections requires accurate sensing of the computational capacity of these heterogeneous chips, the business types suitable for different chips, and their locations in the network, as well as effective management and supervision [222].

(2) Computational power routing based on resource awareness. In computational power networks, after measuring and modeling computational resources, the information is encoded and loaded into network control layer packets for sharing [217, 223]. The network control layer makes network decisions based on shared computational resource information, guiding business routing to different computational resource pools or through collaboration between computational resource pools for business processing, thus enabling network awareness of computational resources to guide global routing.

(3) In-network computing (INC). Leveraging the deployment of programmable network technologies, INC processes packets within the network [224]. Sharing INC power using open and programmable heterogeneous resources accelerates data processing close to the source without altering the original business operating mode, reduces application response delays, and simplifies application deployment processes. Computing power networks support users to dynamically adjust resource scale according to demand to adapt to AI applications of different scales and complexities [223]. In addition, the computing power network can also be combined with other novel technologies. For example, by adopting blockchain technology [225], the computing power network can provide AI with a secure and privacy-protected computing environment, ensuring the transparency and traceability of the computing process [226]. The introduction of technologies, such as smart contracts, can further protect user privacy and data security, allowing users to safely use the computing power provided by the computing power network to process sensitive data and tasks and improving overall security and credibility [227].

In [228], a Kubernete-based prototype testing platform for the computing power network using a microservices architecture was implemented, achieving key enabling technologies for the computing power network, including computational modeling, sense, announcement, and offloading. The authors of [229] described a lightweight multi-cluster hierarchical edge resource scheduling scheme based on a cloud-native resource scheduling mechanism, successfully managing and deploying many heterogeneous edge devices within a unified framework using a lightweight cloud-native platform. An integrated ICT network architecture of "connectivity + computing + intelligence" was proposed in [220], which can sense computational resources and perform related management and control. Numerical results in [230] indicate that on-demand compute resource scheduling scheme significantly improves the efficiency and stability of task offloading and compute resource scheduling in edge computing networks. Detailed power analysis of several INC use cases in [224] showed that INC becomes more energy-efficient at very low processing loads, with increased processing loads having little impact on INC power consumption. In [231], the fog computing network offers powerful computing power support for AI model training, reasoning, and other tasks. This helps shorten model training cycles, improve training efficiency, and enable AI models to iterate and optimize faster. The study presented in [232] takes DNNs as a typical AI application, and

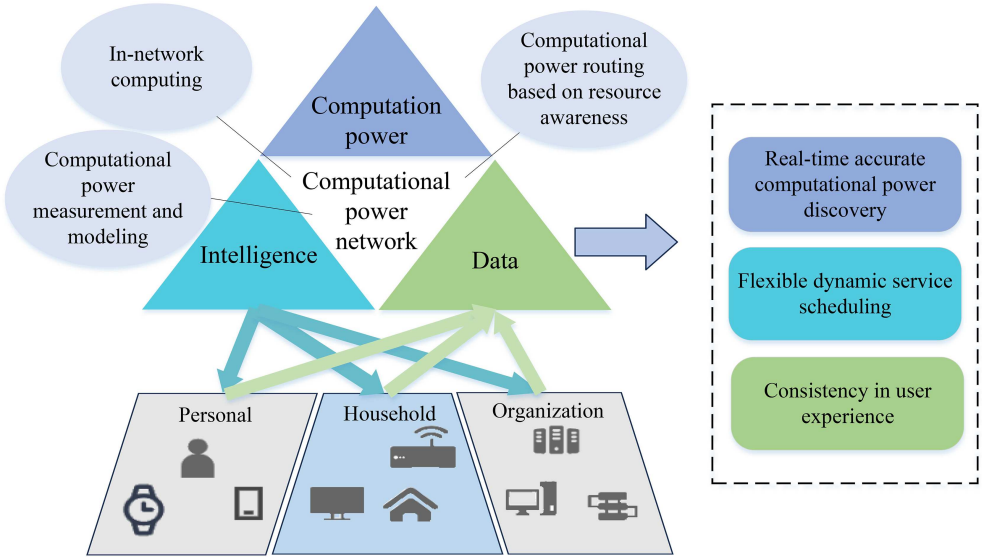

Figure 16 (Color online) Computing power network and AI.

formulates optimization problems under the constraints of energy consumption, delay, computing, and communication resources to efficiently utilize computing resources and avoid resource waste.

3.2.4 *Secure multi-party computation*

In highly complex, heterogeneous, and dynamically evolving network environments, it is imperative to ensure data security while achieving efficient collaboration. As network technology continues to evolve, so do network attack methods, becoming increasingly sophisticated. Conventional security mechanisms are often inadequate to counter these emerging threats. As shown in Figure 17, secure multi-party computation (SMPC) [233,234] is an encryption protocol that allows multiple parties to jointly compute a function while keeping their input data private. Each party obtains the correct computation result without knowing the other parties' private inputs. SMPC has characteristics such as decentralization, input data security, and accurate computation results. It enables AI to handle complex network environments, protect user privacy and data security, support emerging application scenarios, and promote data sharing and utilization.

The main supporting technologies of SMPC include garbled circuits [235], oblivious transfer [236], secret sharing [237], and homomorphic encryption [238].

(1) Garbled circuits technique involves compiling computational logic into a circuit and encrypting each gate (e.g., AND, XOR, etc.) within the circuit. Participants can complete the computation by interacting with the encrypted information without knowing the specific logic of the circuit.

(2) Oblivious transfer is a protocol that allows one party (e.g., the sender) to send one of many pieces of information to another party (the receiver), where the receiver can choose only one piece of information. The sender does not know which piece of information the receiver has chosen.

(3) Secret sharing involves splitting a secret into multiple shares and distributing them among participants. The original secret can only be reconstructed when a sufficient number of shares are combined.

(4) Homomorphic encryption is a special encryption method that allows specific computations (such as addition and multiplication) to be performed on encrypted data. The result of these computations, when decrypted, corresponds to the result of the same computations performed on the original data.

Recent research has leveraged SMPC to achieve secure model aggregation in FL, which involves training ML models on decentralized devices while keeping the training data localized. To counter reverse attacks on data, a privacy-preserving data aggregation mechanism based on secret sharing techniques in SMPC can efficiently protect FL against such attacks [239]. To address high communication overhead and poor scalability in conventional SMPC, a two-phase SMPC-based FL framework has been proposed [240]. This framework enables multiple clients to collaboratively train AI models while protecting their data privacy. Furthermore, a hierarchical FL architecture utilizing SMPC has been developed to alleviate the communication costs and scalability issues associated with aggregating global models at only a few nodes,

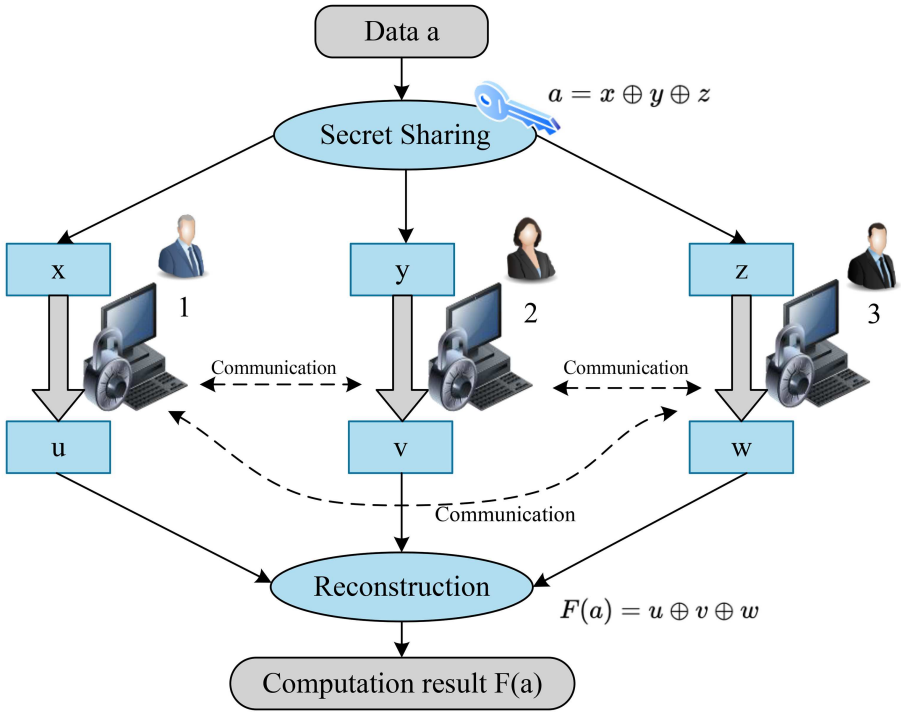

Figure 17 (Color online) Secure multi-party computation [234].

enhancing communication efficiency and scalability without compromising privacy [241]. A novel hybrid FL architecture has been proposed, which combines FL with trusted execution environments, SMPC, and Beidou satellites. This architecture achieves secure key distribution, encryption, and decryption, and provides verification mechanisms for each participant to ensure the security of local data [242]. Additionally, SMPC can combine with blockchain technology to track and mitigate malicious behavior, achieving a secure and trustworthy collaborative edge learning framework [243].

SMPC protects data privacy while enabling collaborative computation across multiple parties, making it a valuable asset for 6G applications. It can contribute to FL, distributed resource allocation and optimization, secure data fusion and analysis, distributed identity verification and key management, and anti-fraud billing systems. By leveraging encrypted computation, SMPC ensures the secure processing of sensitive information. While SMPC enhances data sharing privacy and security in 6G networks, challenges remain, including computational and communication overhead, protocol standardization, and finding a balance between privacy protection and efficiency.

3.2.5 *Semantic communication*

Semantic communication, commonly relying on DNNs, is considered a promising technology in 6G. Compared with traditional syntactic communication which focuses on the accurate transmission of data in bits, semantic communication only transmits the semantics of the message. It can also preserve essential semantic relationships when transmitting information, thereby aiding downstream AI tasks. These characteristics help meet the demands of new data-hungry 6G applications, e.g., holographic communication and XR, as shown in Figure 18.

Figure 19 shows a basic semantic communication model, encompassing three essential components: the semantic encoder, the channel, and the semantic decoder [244]. At the transmitter, the semantic encoder extracts and encodes the SI of the source data. This process involves extracting the SI and compressing or removing the irrelevant information. To achieve this, the raw data are encoded by a neural network, denoted as $\mathcal{F}_\theta$, which outputs the semantics that retain the critical meaning while discarding non-essential details. The compact semantics are then transmitted over a noisy physical channel, such as a wireless channel. At the receiver, the semantic decoder decodes the received data. This process involves "understanding" and inferring the semantic content sent by the transmitter. To accomplish this, the semantic decoder, denoted as $\mathcal{G}_\phi$, processes the received data with the goal of mapping it back to the

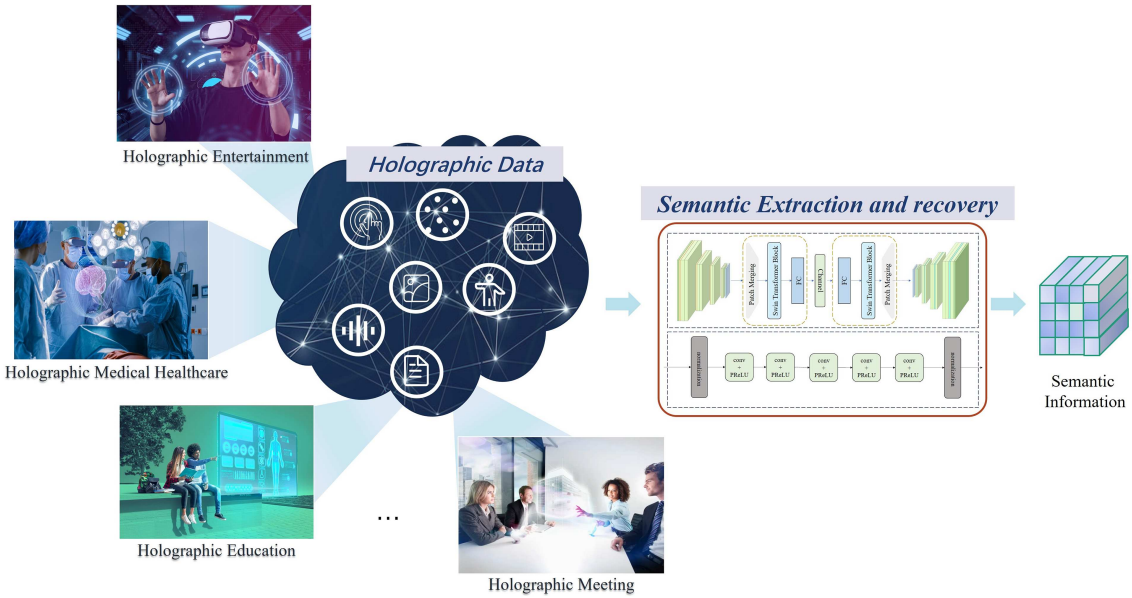

Figure 18 (Color online) Illustration of applications that require ultra-low-latency interaction in 6G networks, and DL-based semantic communications can greatly benefit these applications.

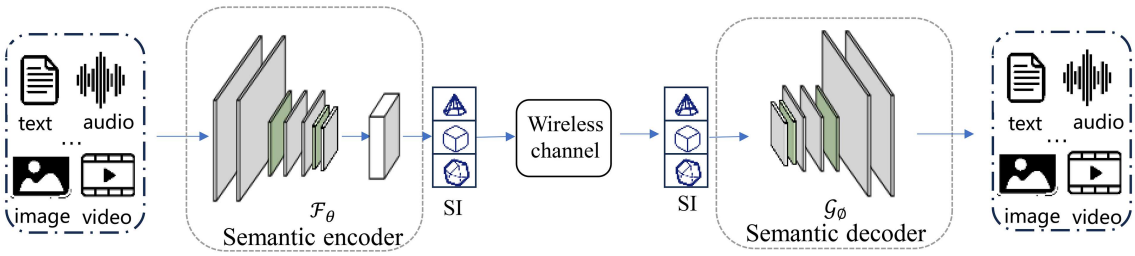

Figure 19 (Color online) Basic semantic communication model.

source data.

The authors of [245] provided a pioneering work in SI theory, establishing the mathematical framework for SI theory. The concept of modulation division multiple access technology was proposed in [246], which efficiently distinguishes users based on semantic features from the information dimension, thereby improving resource utilization in multi-user semantic communication systems. Semantic extraction and recovery, also called semantic encoding and decoding, are the core components of semantic transceivers with four dominant advanced semantic extraction techniques [247]: DL-based semantic extraction, RL-based semantic extraction, KB-assisted semantic extraction, and native semantic extraction. Moreover, most semantic communication studies focus on text transmission, audio transmission, and recognition, image transmission and recognition, and video transmission and recognition from the perspective of data modality and task classification.

(1) Text transmission. The authors of [248] developed a DL-based text semantic communication system named DeepSC, which performs joint semantic-channel coding to achieve superior performance gains compared to existing technologies, particularly at low SNRs. Recognizing that most semantic metrics are non-differentiable, an RL-based optimization paradigm for large-scale and complex text semantic transmission was introduced in [249]. This technique uses self-critical random iterative updates to train decoupled semantic transceivers, addressing the non-differentiable semantic channel optimization issue. The authors of [250] proposed a text-oriented semantic communication technique named semantic coding Reed-Solomon hybrid automatic repeat request (HARQ) by combining HARQ and Reed-Solomon channel coding with semantic coding to tackle inefficiency and inflexibility with varying sentence lengths.

(2) Audio transmission and recognition. A DL-based audio semantic communication system named DeepSC-S was developed, as noted in [251, 252]. The audio semantic communication scheme employs a joint semantic encoder/decoder and channel encoder/decoder to mitigate the semantic dis-

tortion caused by the noises and interference from the wireless channels. The authors of [253] explored FL-based audio semantic communication, developing a wav2vec-based autoencoder composed of CNNs. This system effectively encodes, transmits, and decodes audio SI, reducing communication overhead.

(3) Image transmission and recognition. The authors of [254] investigated a DeepJSCC image semantic communication architecture, where the encoder and decoder functions are parameterized by CNNs and jointly trained on the same dataset to minimize the mean squared error of the reconstructed images. Joint training enables DeepJSCC to avoid the cliff effect and adapt to the varying SNR.

(4) Video transmission and recognition. The authors of [255] developed a variable-length DeepJSCC system that utilizes nonlinear transformations and conditional coding architectures to adaptively extract the semantic features of video frames. This system outperforms the traditional wireless video coding transmission on recognition accuracy.

Semantic communication has unique advantages in semantic understanding, driving advancements in applications such as the tactile internet, holographic communication, XR, and human-machine symbiotic networks, fostering an increasingly intelligent and efficient network.

4 Wireless network large model

LAMs exhibit remarkable capabilities and great potential in academia and industries. However, such a trend presents new challenges for the underlying network infrastructure. Wireless networks play a vital role in the infrastructure and are directly associated with LAMs' performance (e.g., data transmission, processing speed, etc). Therefore, this section aims to explore the critical role of wireless networks in supporting the operation of these LAMs and conduct a systematic analysis and reflection on the relevant issues related to wireless networks in the context of constructing LAMs.

4.1 Comparisons between LLM and wireless network large model

LLMs [256] have demonstrated their promise in text and image tasks. It is challenging to directly apply traditional LLMs to wireless communication systems, as the data in the system consists of various protocol data in access networks, CN, and application data from the upper layer. Furthermore, these data are either structured or non-structured and may be heterogeneous with temporal and relational features. Specifically, the air interface involves wireless channels' transmission and scheduling management, encompassing diverse information generated by various wireless communication standards, such as LTE and 5G. The CN handles routing, authentication, and service control of user data, which includes sensitive content such as user behavior and location information, demanding high levels of privacy and security. Operational management data cover status monitoring, fault diagnosis, and performance optimization of network equipment. These datasets are typically large-scale, high-dimensional, lacking effective designs for unifying data across different standards. These characteristics necessitate the consideration of complex factors such as spatiotemporal relationships, real-time requirements, and system stability when processing and analyzing wireless communication data, presenting significant differences and challenges compared to conventional LLM data processing approaches, as illustrated in Figure 20.

Moreover, wireless communication systems constitute a highly complex ecosystem [257] that employs multiple technological paradigms to meet diverse technical requirements and environmental scenarios. Unlike the relatively lenient requirements for inference speed in LLMs used in text and image processing, the demands placed on model inference speed in wireless communication tasks are critically important. Insufficient inference speed in a wireless network large model can lead to outdated results due to significant changes in channel conditions and signal characteristics during the inference process. Therefore, ensuring that models in wireless communication systems can quickly obtain inference results is a KPI for deployment effectiveness.

On the other hand, the requirements for inference accuracy in wireless network large models far exceed those of LLMs. In language models, different combinations of word arrangements have minimal impact on semantic expression; however, in wireless communication systems, which rely on exact mathematical models, sensitivity to inference errors is markedly high, with repercussions that are difficult to quantify. Signal modulation, a critical process encoding information onto carrier signals, exemplifies this sensitivity. Minor deviations in predicting modulation parameters by the model, such as errors in frequency or phase, can prevent the receiver from correctly demodulating signals, resulting in data loss or errors. In distributed wireless networks, even slight errors in predicting clock offsets or sync signals by the model can

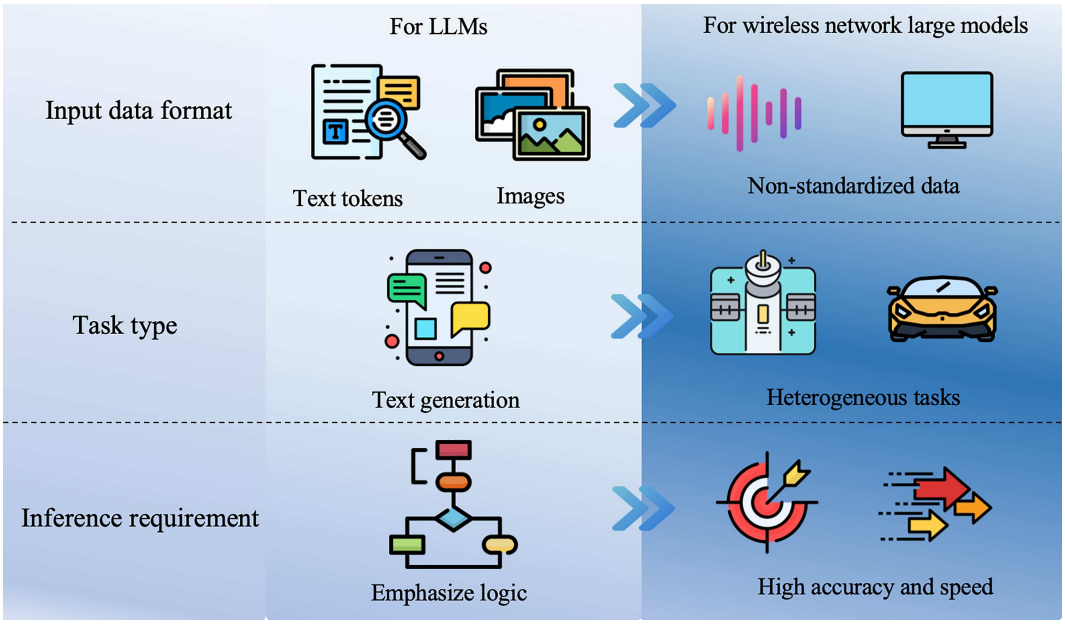

Figure 20 (Color online) Differences between LLMs and wireless network large models.

lead to synchronization failures among nodes, thereby compromising network stability. Therefore, precise control over the inference accuracy of the wireless network large model is a critical research challenge.

Although LLMs have significantly advanced text and image processing, directly applying them to construct a wireless network large model is impractical. In practical applications, integrating domain expertise with advanced data processing techniques is essential for effectively handling and analyzing wireless communication data, addressing its complexities and dynamic challenges.

4.2 Potential approaches to constructing a wireless network large model

Wireless network large model [258] refers to an intelligent and efficient solution to communication network operations, achieved through integrating knowledge and technologies from the communication domain. A wireless network large model provides network management, and operations fault detection and diagnostic services. The failures in networks may lead to performance degradation and service interruptions. By analyzing massive network data, a wireless network large model can monitor network status in a real-time fashion, quickly identify and locate potential reasons, and provide accurate diagnosis strategies.

A wireless network large model optimizes task orchestration and scheduling. Modern communication networks involve complex and diverse tasks and business processes, requiring efficient scheduling and management to ensure network operations. Through an in-depth analysis of network topology and business requirements, a wireless network large model can potentially optimize task execution sequences and resource allocations, enhancing network resource utilization and business processing efficiency.

In terms of performance optimization and resource allocation [259], the wireless network large model employs intelligent strategies and algorithms. By continuously monitoring network load and traffic conditions and integrating predictive analytics and dynamic adjustment mechanisms [260], it achieves real-time optimization of network performance and rational allocation of resources, ensuring network stability and reliability. The prevailing industry viewpoint suggests that the wireless network large model should adopt a three-layer model structure, involving L0, L1, and L2 layers.

- The L2 layer can provide customized solutions tailored to specific business scenarios in this framework. For instance, the L2 layer model can optimize energy utilization efficiency in energy conservation within networks by analyzing data traffic and device loads. Regarding traffic prediction and fault monitoring, the L2 layer model can employ DL techniques to identify anomalies and issue early warnings, thereby enhancing network stability and reliability.
- The L1 layer involves domain-specific grand models. The design of the L1 layer emphasizes modeling and optimization for specific domain challenges and issues. For example, the L1 layer model can analyze wireless channel conditions and user behavior in the air interface to improve data transmission rates

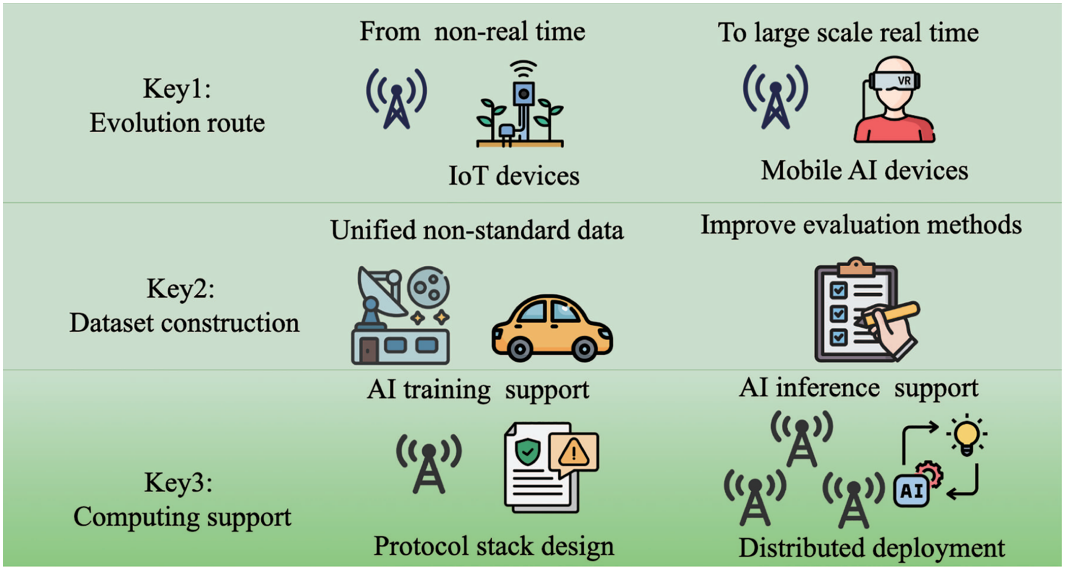

Figure 21 (Color online) Construction philosophy of wireless network large model.

and coverage. The L1 layer model can utilize network topology and device status information for fault diagnosis and optimization scheduling in the CN and operations.

• The L0 layer represents the universal grand model for the entire network. The design of this layer needs to consider the overall architecture and operational mechanisms of communication systems to ensure it possesses strong generalization capabilities. Through comprehensive data analysis and intelligent decision-making, it aims to provide optimization and management services for all communication tasks.

Figure 21 shows the construction philosophy for constructing a wireless network large model in communication networks, encompassing evolution route, dataset construction, and computational support, which complement each other to form the fundamental framework of wireless network large models.

Key 1: evolution route. Constructing a universally applicable large model for the entire network faces many challenges. On the one hand, data within communication networks exhibit highly structured characteristics, encompassing extensive information from various devices and sensors [261], including network traffic, device status, and user behavior. The diversity and complexity of these data present significant challenges for model optimization. On the other hand, tasks within communication networks can be complex and interdependent. For instance, network traffic characteristics may vary substantially across different business scenarios, rendering conventional unified modeling approaches less feasible. A phased approach to model construction becomes important to address the aforementioned challenges effectively. The initial focus can be small-scale, offline L2-level models. Through the practical validation of critical business scenarios such as network energy efficiency, traffic prediction, and fault monitoring, efficient utilization of network resources can be achieved. In designing L1-level models, emphasis should be placed on addressing domain-specific challenges, employing targeted modeling and optimization approaches to achieve domain generalization capabilities across air interface, operations, and CN sides. To construct a universally applicable L0-level model encompassing the entire network, carefully considering the communication system's overall architecture and operational mechanisms is essential to ensure the model's ability to support services and strategies for all communication tasks.

Key 2: dataset construction. Dataset construction is crucial for large-scale network models [262] directly impacting AI model training efficacy and application. Challenges revolve around data acquisition and quality difficulties [263]. To tackle the difficulties in data acquisition, multiple avenues should be explored, including proactive collaboration with industry partners to promote data sharing and openness [264], involving telecommunications operators, equipment manufacturers, and others to expand the coverage of datasets. Furthermore, establishing industry standards and guidelines can guide stakeholders in adhering to unified data collection and processing standards, facilitating data sharing and exchange to overcome data acquisition challenges. A comprehensive data quality assessment framework is also indispensable [265]. Utilizing data quality metrics such as accuracy, completeness, and consistency, data should be assessed and monitored to detect any data quality anomalies and take corresponding measures for improvement and optimization. Employing data quality management tools and technologies

can automate quality checks and validations. During the data collection process, it is crucial to enhance the monitoring and management of data sources to ensure the comprehensiveness and reliability of data acquisition. Advanced data cleaning and preprocessing techniques should be implemented to denoise and handle missing values in raw data. Moreover, establishing robust data standardization procedures to unify data formats reduces dataset heterogeneity and inconsistencies.

Key 3: computational support. Regarding the computational demands of 6G communication to support large-scale network models, optimization efforts are needed in both AI training and deployment. This aims to achieve efficient, reliable, secure, and stable data transmission and utilization of computing resources. In AI training services, key factors include efficient and reliable training data collection and transmission processes, proprietary protocol design, support from network elements, and customized traffic scheduling algorithms [266]. When designing core and access networks, proprietary protocols tailored specifically for AI task transmission should be developed. These protocols need to account for the unique characteristics of AI tasks, such as large data volumes and computational intensity, to ensure optimal transmission efficiency and performance. When designing network elements to support AI task transmission, considerations should be made regarding hardware architecture and software functionality to process and transmit AI training tasks effectively. In traffic scheduling, selecting appropriate strategies is crucial for supporting the training of LAMs. Conventional traffic scheduling strategies may not suffice for AI training tasks, necessitating the design of customized traffic scheduling algorithms tailored to the characteristics of AI tasks. These algorithms should factor in task priorities, network load conditions, and device resource availability to ensure that AI training tasks receive sufficient network and computing resources promptly.

In AI inference services, the high demands of LAMs on storage space and computational capabilities often exceed what a single BS can provide. A viable approach is designing an effective architecture utilizing distributed node collaboration [267] to collectively handle model storage and inference tasks. Alongside the design of distributed node collaboration architecture, effective management and scheduling of inference tasks are crucial to maximize the utilization of each node's computational resources while ensuring system efficiency and stability [268]. A key consideration is the design of task allocation and scheduling algorithms. This approach allows for the rational allocation and scheduling of inference tasks. Moreover, attention should be given to enhancing system monitoring and automated management. Automated management tools and mechanisms, such as automated configuration and deployment, fault diagnosis, and recovery, can be introduced to detect and promptly address faults and anomalies. This reduces management and maintenance costs and improves system maintainability and manageability. These optimization measures collectively provide better technical support and assurance for AI large model computation capabilities in 6G networks.

4.3 Challenges of wireless network large model development

• **Data heterogeneity.** In wireless networks, the diversity and complexity of data present significant challenges in constructing comprehensive large-scale training and testing corpora. The network contains various types of structured and unstructured data, including uplink and downlink channel information, measurement reports, drive test data minimization, KPI monitoring data, network topology performance indicators, and cost and energy consumption logs. These data sources are highly heterogeneous, and the data from different origins often exhibit notable discrepancies in granularity and time resolution. It is necessary to consider how to effectively align these heterogeneous data to build a consistent and representative dataset. Exploring techniques such as data normalization [269], time-series alignment [270], and data fusion [271] can contribute to constructing a high-quality training corpus. Additionally, attention should be given to transforming the processed data into input formats suitable for inference algorithms. This may involve feature extraction and selection, data dimensionality reduction, and the creation of appropriate input representations to enable the model to perform efficient inference and decision-making.

• **Real-time inference.** Wireless network large model can be applied to real-time decision-making tasks such as dynamic spectrum allocation, power control, and user access management. However, excessive inference time can degrade overall network performance and user experience. This is particularly critical in network resource scheduling, where strict time constraints demand that model inference be completed within milliseconds to effectively manage and allocate communication network resources. This directly impacts the real-time responsiveness and service quality of the network. To meet these requirements, the design and deployment of large-scale models in wireless networks must prioritize optimizing

inference speed, which includes hardware acceleration, algorithm optimization, and network architecture adjustments [272]. Furthermore, achieving a balance between model accuracy and computational efficiency is crucial for efficient and reliable network resource management and scheduling.

• **Ambiguous route.** Building a universal large-scale model framework for wireless communication is a challenging task. Current industry practices tend to leverage large-scale pre-trained language models as foundational components, which are then fine-tuned for specific industry applications. However, the complexity and dynamics of wireless networks require models with higher adaptability and flexibility. Scholars have made some progress in exploring models based on semantic communication, which has advanced the understanding and processing of data flows in wireless networks [273]. However, the feasibility and effectiveness of these models in practical deployments, including their stability and performance under varying environments and conditions, remain unresolved and require further empirical validation.

• **Protocol compatibility.** The integration and deployment of the wireless network large model inevitably rely on the communication network's protocol stack. However, since large-scale model technologies have only emerged in recent years, compatibility issues with existing communication protocol stacks remain to be addressed. Whether the current protocol stack architecture can meet the future demands of large-scale models, and how to enhance and optimize the existing protocol stack to better support the application of these models in wireless networks, pose significant technical challenges. This requires an analysis of the existing protocol stack to identify potential barriers to large-scale model integration and the exploration of new protocol design principles to ensure the efficient and stable operation of large-scale models in wireless networks.

• **Security threats.** Wireless network large model faces multiple security challenges [274]. Firstly, it may encounter privacy attacks. Data sets from 6G networks may involve sensitive information, such as user location, mobility patterns, and communication content. Without adequate protection, these data are vulnerable to privacy threats, which can lead to a loss of user trust and legal issues. Secondly, data tampering is another significant threat to large-scale models. Attackers may influence the model's learning process and decision-making outcomes by injecting misleading data or manipulating data sources. For instance, incorrect input data could lead to misjudgments of user demands or network congestion conditions. Additionally, network attacks targeting large-scale wireless network models are also hazardous. Malicious actors may disrupt model operations or regular requests through flooding attacks or resource depletion attacks, severely affecting service availability and real-time decision-making capabilities. Finally, ensuring the security and integrity of large-scale model architectures is critical. Reverse engineering and model parameter leakage could allow attackers to gain sensitive information about intellectual property or operational mechanisms, thereby jeopardizing the stability and reliability of services.

5 AI as a service

To more effectively support "native intelligence" and achieve ubiquitously "universal intelligence", future 6G networks will treat AIaaS, giving rise to the concept of 6G AIaaS [275], which is presented in Figure 22. 6G will provide efficient end-to-end support for AI-related businesses and applications, intelligently connecting distributed agents for large-scale AI deployment across various industries. 6G AI services will address needs for high real-time performance, high security, and privacy, or low overall energy consumption by conducting AI training or inference within the network, offering intelligent capabilities tailored to different application scenarios.

6G networks, as inherently intelligent architectures, will be able to efficiently train or infer large-scale distributed AI through resources and functionalities within the network, including communication, computation, datasets, and foundational models [276]. This transition from AI4NET in 5G to networks for AI in 6G will ultimately provide intelligence that goes beyond data, directly targeting users. This transformation will redefine the ecosystem of edge clouds and create new business models through 6G mobile networks, transitioning from connectivity-focused to service-oriented networks and ultimately achieving ubiquitous intelligence.

5.1 Typical scenarios of AIaaS

Typical scenarios for 6G AIaaS are contexts and situations where AI in 6G networks comes into play, including but not limited to industries and fields such as agriculture, education, and healthcare [135].

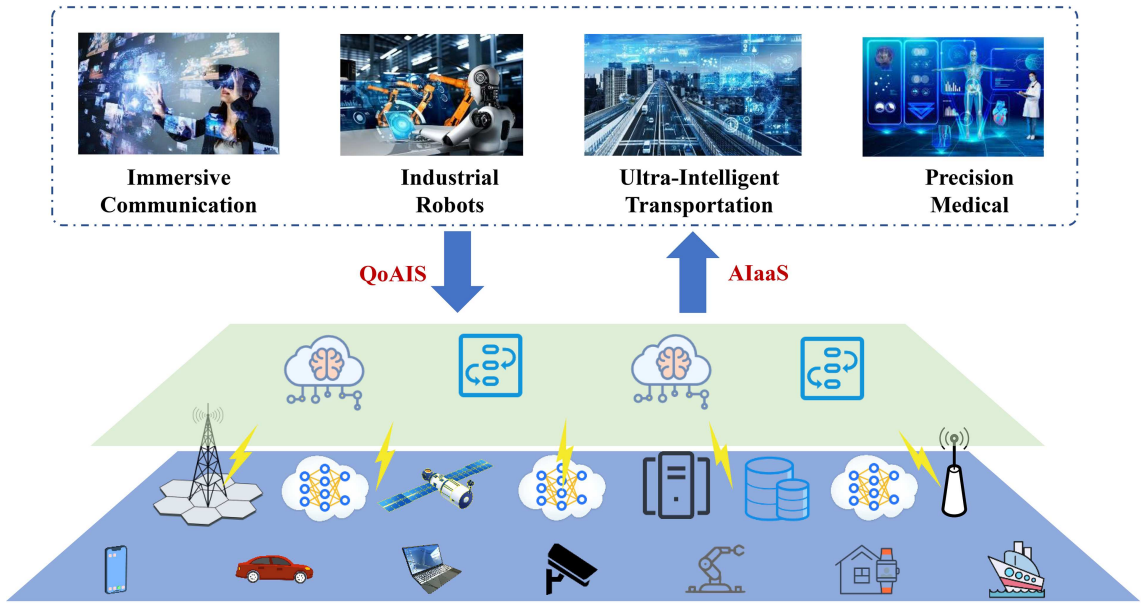

Figure 22 (Color online) AI as a service.

5.1.1 *Immersive communication*

In future communication, immersive technologies like VR, AR, and measurement reports will enhance interactive experiences [277]. AIaaS will support these technologies with real-time language translation, emotion analysis, and personalized suggestions in virtual environments. Trained AI models will be validated to ensure realistic and credible interactions. 6G AIaaS will offer comprehensive AI training, collecting user data for multimodal learning and enabling rapid model training and updates [278]. It will support multi-device linkage, allowing various devices to interact with virtual environments. 6G's extended coverage will seamlessly connect virtual and real worlds, facilitating cross-system information sharing. DT technology will enhance AI validation, generating richer sample data than physical environments and improving model stability and performance. As the number of intelligent terminals increases, 6G's ultra-large connectivity and low latency will be essential for AI training and validation, supporting precise positioning and massive data collection. 6G AIaaS must deliver high-quality autonomous services to ensure secure interactions in virtual environments, with options ranging from fully autonomous to human-controlled services. This support will lead to more realistic and efficient immersive communication, advancing remote collaboration and communication.

5.1.2 *Intelligent industrial robots*

Industrial robots will be widely utilized in future industrial manufacturing and transportation scenarios [279]. 6G AIaaS provides AI training services for robots, such as data collection through robots, applying model training for multi-agent learning, and distributing the trained models to the robots. The extreme transmission capabilities of 6G will support rapid model training and parameter exchange between robots and the network. 6G AIaaS also offers multi-system linkage capabilities, allowing various terminal devices to network with robots flexibly and enabling remote control via computers, smartphones, and VR devices [280]. Robots can connect and interact with the environment through 6G AIaaS's sensing technology. They can even engage in cross-environment and cross-system learning with robots from different departments, sharing experiences. Additionally, the 6G network can leverage DT networks to provide AI verification services, iteratively training AI models with greater robustness and better performance through performance pre-validation. 6G AIaaS offers AI verification services for trained models, requiring the DT network to generate sample data in more scenarios than the physical environment, reducing the overhead and performance impact of data collection from the physical network.

5.1.3 *Precision medical*

Smart healthcare will encompass various aspects of disease prevention, prediction, diagnosis, reasoning, monitoring, clinical surgery, patient care, and vaccine and drug development throughout the entire lifecycle [281]. The new 6G system will better support the massive information transmission and synchronization required for smart healthcare and directly empower the processing and decision-making of medical information. By leveraging AI over 6G networks, geographically dispersed medical institutions, and individual practitioners can connect more extensively, aggregate more AI models and case data samples, and rapidly transmit and synchronize information between doctors and patients. This enhances FL and group learning, continually improving the accuracy, reliability, and real-time performance of predictive diagnostics and treatment actions. Smart healthcare can also achieve more rational planning and optimized allocation of various medical resources through 6G network AI, significantly reducing the physical and psychological stress on healthcare practitioners and patients.

5.1.4 *Ultra-intelligent transportation*

Ultra-intelligent transportation will involve various stages across the temporal dimension, including high-definition map downloads, vehicle environment sensing, environmental prediction, and route planning [282]. Taking autonomous driving as an example, the 6G system can provide AI-based data services, AI computation offloading services, and AI-based environment prediction or path planning. AI-based data services refer to the 6G system providing autonomous vehicles with AI-driven environmental sensing results. For instance, when an autonomous vehicle encounters a sensing blind spot, the 6G system can collect sensing data through sensors or wireless signals, then use AI models to infer the environment within the blind spot and feed the results back to the vehicle, enhancing driving safety. AI computation offloading refers to the process where autonomous vehicles offload part of their AI model inference or training computations to the 6G system. For example, when a vehicle drives through a complex road segment, the computational power required for AI model inference can increase significantly, resulting in unacceptable latency. In such cases, the vehicle can offload some of the AI computation tasks to the 6G system, thereby reducing the processing burden on the vehicle's chips.

With the 6G network's DL and multi-sensor fusion capabilities, an integrated sensing and decision-making architecture is established for vehicles, roadside infrastructure, and cloud platforms [283]. This architecture advances a new era of intelligent transportation characterized by intelligent travel, ubiquitous services, and comprehensive management and control.

5.2 QoAIS

Quality of AI service (QoAIS) refers to a comprehensive metrics framework for evaluating AI services within a network. It encompasses key dimensions such as performance, connectivity, computation, data, security, and orchestration [43, 275, 276, 284]. Notably, the performance metrics in QoAIS differ from traditional model KPIs used in ML. While model KPIs primarily assess the internal evaluation metrics of a model-serving purposes like optimization and fine-tuning-QoAIS extends this scope. It incorporates factors such as the generalizability and robustness of models or algorithms, thereby offering a holistic view of system-wide performance. The metrics for other dimensions within QoAIS are novel, as they specifically account for the unique characteristics and requirements of 6G networks. As shown in Table 3, the QoAIS system should include as following.

• **Performance-related indicators.** Aspects critical to evaluating and improving AI models are encompassed, including the boundary of performance metrics, training time, generalization, reusability, robustness, interpretability, and consistency between the loss function and optimization objectives. Boundary of performance metrics defines the upper and lower limits for assessing the quality of model performance indicators, such as error rate, precision, and recall. Generalization reflects a model's ability to make accurate predictions on new, unseen data, while reusability highlights its capability to remain effective across diverse scenarios. Robustness ensures the model maintains consistent performance despite perturbations, adversarial attacks, or input uncertainties. Interpretability focuses on the extent to which the model's internal workings and outputs can be understood and explained, fostering transparency and trust. Consistency between the loss function and optimization objectives ensures alignment between the design of the loss function during training and the AI system's goals, accounting for all relevant variables to achieve the desired outcomes.

Table 3 QoAIS indicator system of AI services.

Indicator dimension	QoAIS indicators
Performance	Boundary of performance metrics, training time, generalization, reusability, robustness, interpretability, consistency between loss function and optimization objectives, fairness, etc.
Connection	Bandwidth and jitter, link delay and jitter, bit error rate and jitter, reliability, etc.
Data	Feature redundancy, completeness, data accuracy, time-consuming data preparation, sample space balance, sample distribution dynamics, etc.
Computation	Computing accuracy, duration, efficiency, etc.
Security	Information confidentiality, data/algorithm privacy levels, data authenticity, traceability, etc.
Orchestration	Full autonomy, partially controllable by humans, fully controllable by humans, etc.

• **Connection-related indicators.** Reliable and efficient AI services depend heavily on network characteristics, including bandwidth, jitter, link delay, bit error rate, and overall connection reliability. Bandwidth represents the data transmission capacity of the network, directly impacting the speed and quality of real-time AI applications, such as video analysis or interactive systems. Jitter refers to the variability in packet arrival times, which can lead to inconsistencies in service performance, particularly in latency-sensitive tasks. Link delay quantifies the time taken for data to travel across the network, with lower delays contributing to smoother and more responsive AI systems. Bit error rate measures the rate at which errors occur during data transmission, emphasizing the importance of accurate and reliable data exchange. These factors collectively define the stability and dependability of the connection, directly influencing the QoAISs.

• **Data-related indicators.** The quality and characteristics of data play a vital role in ensuring the success of AI systems, focusing on feature redundancy, completeness, accuracy, and preparation time. Feature redundancy highlights the presence of overlapping or irrelevant features, which can complicate model training and reduce efficiency. Completeness ensures that all necessary data attributes are available to achieve comprehensive insights and predictions. Data accuracy emphasizes the importance of reliable, error-free data to maintain the validity of AI outputs. Time-consuming data preparation reflects the challenges involved in cleaning, transforming, and organizing data for training purposes. Sample space balance ensures equitable representation of different classes or categories within the dataset, critical for minimizing bias. Sample distribution dynamics track changes in data distributions over time, enabling models to adapt and remain effective in evolving environments. Together, these aspects underpin the foundation for building robust and trustworthy AI models.

• **Computation-related indicators.** Efficiency and accuracy in computation are crucial for ensuring reliable and high-performing AI systems, encompassing factors such as computing accuracy, duration, and efficiency. Computing accuracy reflects the precision of the computational results, ensuring the reliability of outputs across tasks. This dimension emphasizes the efficiency and accuracy of AI computations, which include computing accuracy, duration, and efficiency. Duration measures the time taken to complete computational processes, with faster computations enabling real-time or near-real-time performance in AI systems. Efficiency evaluates the resource utilization of the computation, including memory, processing power, and energy consumption. Optimizing these aspects ensures that AI services are not only accurate but also cost-effective and sustainable.

• **Security-related indicators.** Protecting data, algorithms, and systems is a fundamental requirement for AI services, with key considerations including confidentiality, privacy, authenticity, and traceability. Information confidentiality ensures that sensitive data remains inaccessible to unauthorized entities, safeguarding user trust and compliance with regulations. Data/algorithm privacy levels highlight the measures taken to protect proprietary algorithms and personal data, reducing the risk of exposure or misuse. Data authenticity guarantees that the data used in AI processes is genuine and unaltered, ensuring reliable outputs. Traceability refers to the ability to track the origin, flow, and modifications of data and algorithms, enabling accountability and transparency in AI systems. Collectively, these factors contribute to the safe and ethical deployment of AI services.

• **Orchestration-related indicators.** The level of automation and human involvement in AI services plays a key role in determining how workflows are managed and controlled. Autonomy in AI services defines the degree of automation and the extent of human involvement required across data, training,

validation, and inference workflows. It reflects user expectations regarding the level of automation and is categorized into three levels. At the highest level, full autonomy enables AI services to operate independently without human intervention. In contrast, partial human control involves workflows where some stages are automated, but others require human assistance. Finally, full human control necessitates human involvement at every stage of the workflow, with no reliance on automation.

5.3 Key capabilities for supporting AIaaS

The concept of AIaaS envisions a network-wide platform that seamlessly delivers AI services to users and devices connected. Achieving this vision hinges on several critical capabilities, which can be grouped into three primary aspects.

5.3.1 *Data support and model accessibility*

A fundamental requirement for an AIaaS platform is robust data support. This can be achieved by utilizing the network's sensing capabilities and the data collection functions of IoT devices and user terminals to enable real-time data acquisition and storage. Ensuring data privacy while maintaining a direct link between users and their data is crucial, as this connection underpins model training and inference services. Moreover, AI services also rely on diverse models tailored to different tasks. The platform should allow users to access pre-existing models or create custom models to suit their needs. A dedicated interface connecting users to these models would facilitate customization and enhance the flexibility of AI services.

5.3.2 *Distributed computing and resource optimization*

The distributed computing power inherent in 6G networks-spanning CPUs, GPUs, and cache resources across CN nodes, BSs, user devices, and third-party systems-forms the backbone of AIaaS. By leveraging software defined network, these resources can be discovered, tagged, and virtualized for unified management [285]. Protocols like distributed Hash tables can be employed to store, retrieve, and update resource information efficiently. With resource visibility in place, computing, caching, and bandwidth resources can be allocated dynamically to meet service demands. Techniques such as DRL, game theory, and model predictive control enable adaptive resource allocation and service scheduling, ensuring efficient operation even under varying workloads.

5.3.3 *Security and lifecycle safeguards*

The distribution of cloud-edge computing nodes introduces new challenges in securing dynamic and cross-domain network connections. Comprehensive security measures are needed to protect AI services throughout their lifecycle. These include protocols for secure data acquisition and device authentication to ensure the integrity of inputs. Secure data fusion algorithms can validate and clean data from multiple devices, while homomorphic encryption offers privacy-preserving data aggregation. Safeguarding models, which represent significant intellectual property, is equally important. Measures like access control, differential privacy, FL, and real-time monitoring of data flow can protect against unauthorized use, model inversion, and other emerging threats.

Integrating these capabilities into AIaaS platforms requires a cohesive approach that addresses data management, computational scheduling, and security concerns simultaneously. The multidimensional service experience provided by these platforms should align with user-specific needs. With the evolution of 6G networks, traditional network limitations will be overcome, enabling AIaaS to thrive within an intelligent, adaptive, and efficient network architecture.

6 Standardization process

As wireless communication and AI integration is a focus of the industry's research direction, important standardization organizations and industry alliances at home and abroad have launched research on integrating mobile communication networks and AI since 2017. These include the 3GPP, ITU, European Telecommunication Standardization Institute (ETSI), China Communication Standardization Association, IMT-2020 (5G) Promotion Group, and IMT-2030 (6G) Promotion Group. Currently, 3GPP has

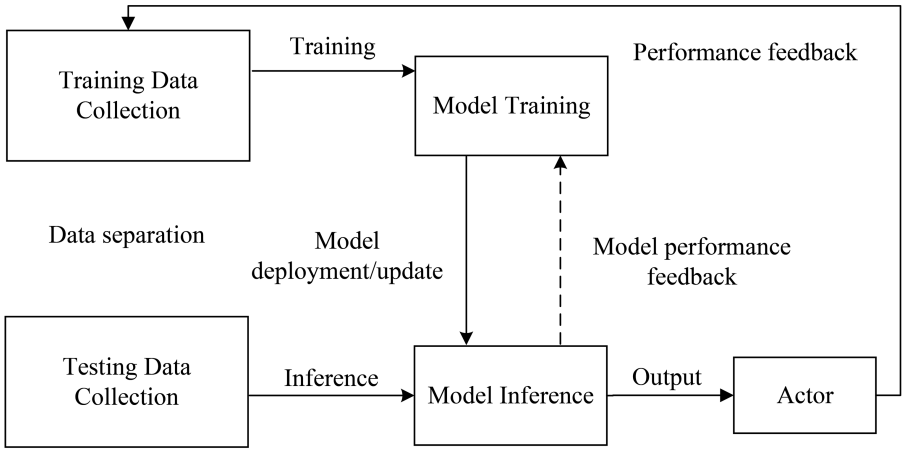

Figure 23 This is the flowchart when applying the AI model to the communication process. The training task of the AI model needs to collect a large amount of data information in advance. The trained model must also be deployed in the corresponding position, and the model inference performance determines whether the AI model still has application value.

three technical specification groups (TSG): the CN, RAN, and service and system aspects (SA). Each TSG has 4 to 6 WG. The 3GPP SA WG2 (responsible for developing the overall 3GPP system architecture and services, including user equipment, access network, CN and terminals (CT), and internet protocol (IP) multimedia subsystem system architecture and services) and the 3GPP RAN WG3 (responsible for the overall radio network architecture and the specification of protocols for the related network interfaces) already accomplished some specification work on data collection, data analytic, and associated procedures at CN and radio network. 3GPP SA WG5 has progressed in standardizing network management data analytics services for managing AI/ML Models and functions. ETSI's zero-touch network and service management WG focuses on the research and analysis of technologies related to automated closed-loop operations, aiming to achieve a high degree of automation in network and service management. ETSI's experiential networked intelligence (ENI) WG is focused on defining an experiential and aware network management architecture that uses AI to improve operators' experience in network deployment and operations. Its goal is to use AI technology to improve the efficiency and automation of network management by defining a perception-adaptive-decision-execute control model. Figure 23 displays the application process of the AI model.

6.1 3GPP

3GPP has conducted relevant research on network intelligence and automation and introduced NWDAF as a CN AI element since Rel-15. NWDAF can collect data from any network functions in 5GC and third-party application functions. Due to time limitations, only slice-specific network data analytics is supported in Rel-15. Various network data and third-party application data from AF and 5GC network functions could be used as input to NWDAF, and output analytic results from NWDAF could be used by other NFs to guide network optimization, such as the following use cases been introduced in the Rel-16: customized mobile management, 5G QoS enhancement, UPF selection, network functions load balancing, and slicing SLA guarantee. Rel-17 allows multiple NWDAFs to be deployed in a distributed architecture. Another enhancement is decomposing the monolithic NWDAF into a model training logic function, analytics logical function, and analytics data repository function. Further, the data collection coordination function enhances the data collection framework. Rel-18, the model performance monitoring, multi-vendor ML model sharing, horizontal FL, and six more use cases have been standardized.

3GPP has identified typical application scenarios for self-organizing networks (SON) and has applied ML techniques, successfully verifying its usefulness in optimizing network performance. In 2018, 3GPP SA5 Group set up a research project, "A study on SON for 5G", aimed at the integration of SON and 5G network technology, studying to promote the integration of SON and 5G network infrastructure by strengthening the collection and application of wireless big data.

The RAN1 WG initiated the study item "Study on AI/ML for NR air interface" as part of the Rel-18 phase [286]. The study item explores how AI/ML algorithms can enhance air interface performance (throughput, robustness, accuracy, or reliability) and reduce overhead. The main application scenarios of

this research include AI-based CSI feedback, AI-based beam management, AI-based positioning enhancement, and life cycle management of AI/ML model/functionality for the air interface. Similarly, the 3GPP RAN2 WG launched the "Study on AI/ML for mobility in NR" research project in December 2023 [287]. The study project evaluates the potential benefits and specification impacts of AI/ML-aided mobility for network-triggered L3-based handoff. The project addresses several critical aspects, including AI/ML-based radio resource management measurement and event prediction for L3 mobility, handoff/radio link failure prediction, and measurement events prediction.

The 3GPP RAN3 WG identified the study item "Study on further enhancement for data collection" in July 2020 [288]. This study item aimed to enable an intelligent functional architecture for RAN. By leveraging use cases, it targeted to enhance data collection and identify potential standardization implications for NG-RAN nodes and interfaces. In February 2022, the study item was completed, and the intelligent typical functional architecture on the RAN side, input, output, and feedback. The potential standard impact of three high-priority use cases (e.g., network energy saving, load balancing, and mobility optimization) was incorporated into TR 37.817. The intelligent universal functional architecture on the RAN side includes functions, such as data collection, model training, model inference, and action execution. Recently, the 3GPP RAN3 WG has launched the "Study on enhancements for AI/ML for NG-RAN" work item [289]. The WI aims to incorporate data collection enhancements and signaling support specified in existing NG-RAN interfaces and architectures to enable AI/ML-based network energy saving, load balancing, and mobility optimization.

The TSG SA (responsible for the overall architecture and service capabilities of 3GPP systems and the cross-3GPP TSG coordination) launched the study item, "3GPP AI/ML consistency alignment", in June 2024 [290]. The study item investigates ongoing AI/ML work in TSG CT, TSG RAN, and TSG SA WGs, and identifies the instances of any potential misalignment and/or inconsistencies. TSG SA leads the close collaboration and takes inputs from TSG CT and TSG RAN. The technical report of this study item will be finalized by June 2025.

6.2 5G-MoNAch

The 5G-mobile network architecture (5G-MoNArch) project in Europe introduced RAN-DAF, an independent AI analysis function explicitly designed for control unit plane in 5G NR, for data analysis and decision-making [291]. As a network element for AI and data analysis on the RAN side, RAN-DAF can monitor and collect UE and RAN data, and AI can prioritize local processing based on these real-time data to solve the need for rapid response of operations such as wireless resource management. MoNArch recommends that the RAN-DAF pass information to the controller on the RAN side to collaborate with the RCA to optimize wireless side performance, such as slice-sensitive selection, live wireless resource control, and cross-slice resource management.

6.3 ITU-T

ITU-T Study Group 13 established the ITU-T focus group on ML for future networks and 5G in November 2017. The group has drafted several technical reports and specifications for ML in future networks, including network architectures, interfaces, protocols, and algorithms. There are three WG under the focus group where WG1 studies potential ML use cases in future networks and identifies the needs of use cases, and WG2 mainly works to classify ML in mobile communication networks and define the required data formats and related mechanisms to protect security and privacy. WG3 primarily studies the requirements of ML on network architecture in mobile communication networks, including network functions, interfaces, and resources.

6.4 ETSI

In 2017, ETSI established the ENI WG dedicated to applying AI technologies to network operations. Improve the experience of operator network deployment and operations with a closed-loop AI mechanism based on context-aware and data-driven strategies. Since its inception, ENI has published several versions of specifications and reports, including system architectures with context-aware policy management, data processing mechanisms, and hierarchical evaluation methods for network AI. In 2019, ETSI ENI published the first edition of this research report and standard and launched the second phase of the closed-loop control study for real-time networks.

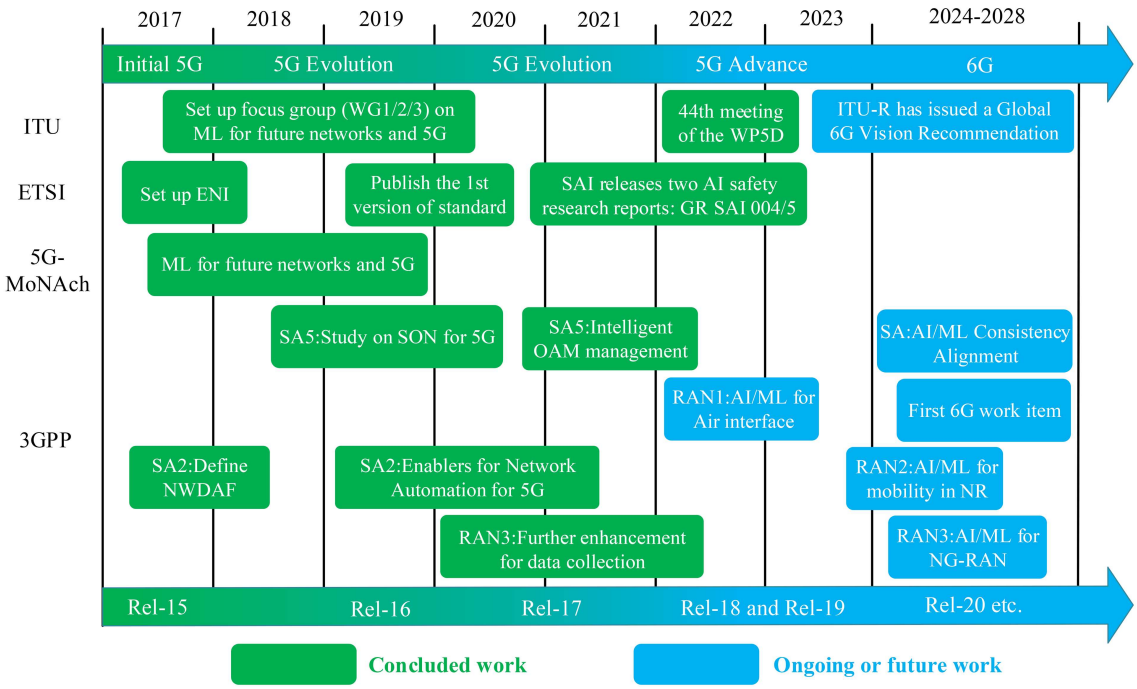

Figure 24 (Color online) Standardization process of AI/ML for wireless networks.

6.5 Collaborative efforts in 6G

In recent years, standardization organizations have also focused on the research and preparation of 6G. In June 2019, the China IMT-2030 (6G) Promotion Group was established to gather the strength of Chinese industry, universities, and research institutes, promoting China's 6G mobile communication technology research and carrying out international exchanges and cooperation and successively releasing several white papers and research reports such as the data format and model proposal of mobile communication and AI integration. In June 2023, the 44th Meeting of the ITU-R Working Party 5D was held in Geneva, Switzerland, and the recommendation on the framework and overall objectives of the IMT for 2030 and Beyond was completed. This proposal aims to set a framework and overall goals for the development of IMT in 2030 and beyond, involving a global 6G vision consensus, 6G goals and trends, typical 6G scenarios, and capability indicators system, including the convergence of AI and communication scenarios, and define AI-related indicators. On September 12, 2024, at the 105th Plenary Meeting held in Melbourne, Australia, 3GPP officially launched the first 6G study item — "6G use cases and service requirements". This initiative formally marks the 6G technology from the pre-research stage into the substantive standardization stage. The vigorous development of AI standardization will significantly promote the integration of communication and AI, as shown in Figure 24. AI still has a way to go from standardization to landing, which is also the main research direction of many standardization organizations in the future.

7 Challenges of AI and communication for 6G

This section delves into multiple challenges confronted by AI and communication in 6G. Herein, AI4NET and NET4AI represent different perspectives of AI empowering networks and networks supporting AI [292], respectively. Sustainability cannot be overlooked in the integration process of 6G and AI. The demands for real-time AI, security and privacy, and human-AI collaboration also bring challenges to the development of AI and communication in the 6G era.

7.1 Challenges of AI4NET

• **Reliability of AI.** The reliability of AI in networks poses a significant challenge due to the dynamic, heterogeneous, and often unpredictable nature of 6G communication systems [293]. AI models, which

rely heavily on data-driven learning, can struggle to maintain consistent performance when exposed to situations that deviate from their training environments. For instance, unexpected network anomalies, rare edge cases, or sudden shifts in user behavior patterns can lead to errors or even system failures [294]. Additionally, the distributed and decentralized nature of 6G networks introduces further complexity, as AI models must operate across diverse nodes with varying resource constraints and environmental conditions. This interconnectedness increases the risk of cascading failures, where an error in part of the system can propagate and disrupt the entire network. An approach based on model-driven engineering principles was proposed in [295], which allows ML experts to schedule the execution of drift-detecting algorithms on a computing cluster. The author of [296] proposed a distributed parallel modeling and monitoring framework for plant-wide processes with big data, where the multilevel monitoring indexes and fault contribution indexes are established based on the Bayesian fusion algorithm.

• **Stability of AI.** Stability refers to an AI system's ability to produce reliable outputs when exposed to variations in input data, environmental conditions, or system parameters. The stability of AI in networks is a critical challenge, as it directly impacts the consistency and predictability of AI-driven decisions in dynamic and complex environments [297]. This is particularly challenging in networked systems due to the heterogeneity and unpredictability of real-world conditions [298]. For instance, fluctuating network traffic, sudden surges in user demand, or unexpected hardware failures can introduce variations destabilizing AI models. The simulation results of [299] indicate that reducing the number of model parameters can enhance the model's robustness against maximum adversarial attacks, while increasing the number of model parameters can improve the model's robustness against minimum adversarial attacks. The authors of [300] studied the robustness of mainstream link prediction methods under various network attacks and discovered that the link prediction method with high performance probably has low attack robustness.

• **Generalization of AI models.** The generalization problem is widely prevalent in integrating AI with communication systems [301]. The application of AI in wireless communications imposes even higher demands on model generalization [302]. In wireless environments, users are situated in varying contexts, with their relative position to the BSs and mobility playing significant roles. This variability means that pre-trained AI models may lose effectiveness when confronted with new data that deviate from their training sets. In the dynamic and complex wireless environments of 6G networks, models with poor generalization capabilities will lack the potential for deployment and widespread use. Data enhancement or meta-learning can potentially improve model generalization. The scenario-adaptive meta-learning for mmWave beam alignment method was proposed to improve the domain adaption capabilities of beam alignment problem in [303]. A fast adaptive channel prediction meta-learning technique combined with a denoising process was proposed in [304] to improve the generalization of the predictor.

• **Interpretability of AI.** DL is often treated as a black box due to its highly nonlinear nature. Humans can hardly understand the meaning, importance, and fluctuation range of parameters in the neural network through basic statistical assumptions like linear regression parameters, leading to unexplainability. Improving the explainability of AI helps humans understand the decision-making process, working principle, and potential bias of AI, and build trustworthy AI [305,306]. However, millions or even billions of parameters in DNNs and their nonlinear activation function form a highly complex decision boundary, hindering the interpretability of a model's decision-making process. Through the gradient derivation and visualization of the model training process, the interpretability of the model can be effectively improved. The authors of [307] deduced the gradient expression of the meta-model to improve the interpretability in the process of backpropagation and promoted the application of meta-learning. t-SNE visualization was proposed in [308] to visualize the representations of data points. t-SNE is an embedded model that can map data from a high-dimensional space to a low-dimensional space and retain the local characteristics of datasets.

7.2 Challenges of NET4AI

• **Dynamics of networks.** Constructing an AI execution environment in wireless networks inherently entails significant dynamic variations. The interference and loss in the wireless connections between terminals and BSs are impacted by the surrounding environment, changes in the operational states of other BSs, and the mobility of terminals. These variations contribute to the dynamic nature of mobile wireless networks, resulting in slow model training convergence and adversely affecting model inference performance, and even causing model divergence. A mitigation strategy for unreliable communication nodes was designed in [162] to prevent unreliable nodes from contributing to stable convergence during

the AI training process. In response to the transmission errors caused by the dynamics of networks, the authors of [309] modified the global aggregation method to counteract the model drift resulting from data packet errors. The theoretical relationship between dynamic unreliable communication channels and collaborative learning was also established in [310].

• **Heterogeneity of networks.** There is a myriad of heterogeneities in wireless networks: communication heterogeneity arising from the geographical locations of nodes and the time-varying nature of communication links; computational heterogeneity due to differences in the number, manufacturing processes, and architectures of node chips; data heterogeneity in the volume and distribution of data possessed by nodes; and model heterogeneity stemming from varying task objectives. Data heterogeneity results in disparities among local gradients computed by AI models. Differences in computational power and communication capabilities lead to variations in model training and transmission efficiency, giving rise to the straggler effect. In [163], the sampling of training nodes in collaborative learning was optimized from the dimensions of communication, data, and computation, achieving efficient training on heterogeneous edge networks. For the heterogeneity of different models, the authors of [311] provided approximate rules for the decision boundaries of each model to bridge the gap caused by model heterogeneity, achieving more accurate model precision.

• **Complexity of networks.** The large scale and complexity of 6G networks significantly affect the deployment and performance of AI systems. In some cases, AI models must train and infer in real time to make effective decisions, which is a daunting task due to the high dimension and dynamic nature of the network data. The deployment of AI in 6G networks also necessitates decentralized architectures, complicating data aggregation and synchronization. Furthermore, the intricate interplay between various network components, such as edge and CN, requires AI models to handle multifaceted interactions and dependencies. To address these challenges, there is a need to advance research in scalable AI techniques. The authors of [312] proposed a hierarchical RL framework based on human prior knowledge. Through process decomposition, stage transformation, key feature selection, and a policy gradient with a parameter-based exploration method, it provides insights in solving complex operations. In response to the differentiated data distribution, a clustering collaborative learning framework was proposed in [313], where the server dynamically determines the optimal number of clusters by iteratively performing incremental clustering. Another focus could be on implementing scalable AI solutions that can dynamically adapt to the network complexity and scale. The authors of [314] used MARL to learn an adaptive fully distributed collaboration strategy for each collaborative node in complex wireless networks.

• **Resource scarcity of networks.** The localized model training and transmission may not always succeed due to constraints on mobile devices' power and computational resources, bandwidth limitations, and wireless channel impairments, directly compromising the performance of FL regarding model accuracy and convergence speed. Consequently, evaluating the trade-off relationship between cognitive performance and multi-dimensional resource allocation becomes imperative upon introducing distributed learning. For FL-enabled edge intelligence, the authors of [315] derived mathematical relationships between FL performance, consumed computational resources, and communication resources, and provided a theoretical foundation for the design of intelligent access networks. To meet the low-latency requirements of multimedia MEC services, a task offloading and resource allocation algorithm based on the double DQN was proposed in [316], which assigns appropriate task volumes to different devices and optimizes their communication and computing resource settings via BS. Moreover, the authors of [317] introduced a distributed resource allocation method. Leveraging meta-federated RL, each device can autonomously optimize its transmission power and channel usage according to wireless environments.

7.3 Consideration of sustainability

While the advanced capabilities of AI are crucial for managing complex 6G operations to achieve unprecedented speeds, low latency, and high reliability, the computational power required comes at a high energy cost, which, if unchecked, could increase carbon emissions. The energy consumption of AI pertains to communication networks, cloud computing centers, and edge computing devices. For example, the computational requirements of LLM (e.g., GPT-4) and complex AI tasks (e.g., RL/DRL) have increased substantially [318]. The training stage may consume several hundred to thousands of MWh of energy, while the energy consumption in the inference stage is directly related to the invocation frequency, model scale, and real-time requirements [319]. For the energy consumption assessment of cloud computing centers and edge, a more comprehensive perspective is needed. For instance, the energy efficiency of

a data center can be evaluated by the power usage effectiveness, which is the ratio of the total energy consumption of the data center to the energy consumption of IT equipment. The cost of leasing these computing resources is affected by the duration of resource usage, the required level of computing power, and market price fluctuations, and is ultimately passed on to end — users in the form of service pricing, subscription fees, etc.

Techniques such as pruning, quantization, and using neural architecture search can help create models that maintain high accuracy while reducing computational load [320]. Energy-aware AI models, which are capable of dynamically adjusting their energy consumption in response to network demands, can also contribute to sustainable operations [321]. Additionally, optimizing data processing and transmission within 6G networks represents another crucial aspect [322]. Integrating renewable energy sources into 6G infrastructure can also contribute to the green 6G vision.

7.4 Requirement of real-time AI

The 6G aims to achieve uRLLC in support of real-time applications, including autonomous driving, telemedicine, and industrial automation. These applications require data transmission and processing to be completed within milliseconds or even microseconds. However, the state-of-the-art AI models, characterized by large parameter spaces and intensive computation, typically employ distributed deployment and incur additional latency [323]. In scenarios such as sensor data fusion for industrial automation and traffic monitoring and analysis for intelligent transportation systems, algorithms are required to process data and make accurate fusion decisions within an extremely short time [324].

DL algorithms construct DNNs to extract deep feature representations from vast and complex data, unlike traditional methods that rely on manually designed feature extraction rules. This significantly reduces errors and limitations caused by human factors, thereby enhancing the accuracy and real-time performance of data fusion. RL/DRL dynamically adjusts fusion weights and methods based on real-time feedback information from different data sources, adapting to the constantly changing data environment and task requirements. Reducing the computational requirements of models without significantly compromising accuracy is possible through designing more efficient network architectures, e.g., lightweight versions of CNN and RNN [325]. Adopting edge computing can reduce latency by shifting data processing tasks from the central cloud to network edge [326]. Innovations in network architecture, e.g., network slicing [327], can also provide customized network resources for different services and applications, ensuring critical tasks receive the necessary bandwidth and latency guarantees. Additionally, dedicated AI processors, such as GPU, tensor processing units, and field-programmable gate arrays, can expedite model inference, helping meet the stringent real-time requirements of 6G networks.

7.5 Security and privacy

AI models are susceptible to adversarial attacks, and imperceptible perturbations in the input data can lead to incorrect outputs [328]. Moreover, the security and privacy risks faced by AI algorithms in the process of data fusion are formidable. Specifically, the data generally originate from diverse data sources, which frequently encompass sensitive information, such as personal identities, medical records, and financial transactions. This gives rise to risks on all aspects of data fusion, ranging from storage and transmission to processing [329, 330]. For instance, the server storing the fused data might be subject to hacker attacks. During the data preprocessing of AI model training, data are susceptible to theft if appropriate encryption protection measures are not in place.

One defense approach is developing robust AI through adversarial training, which can enhance the resilience of AI systems [331]. Another approach is FL and/or MARL, where AI models are trained across multiple devices without exchanging the raw data, thereby preserving privacy [332,333]. For example, the authors of [334] proposed an adaptive feature-correlation region segmentation mechanism to offer privacy protection for DL models under a given moderate privacy budget. SMPC and homomorphic encryption can also be performed to ensure data privacy even when some processing needs to be conducted at third parties [335]. Moreover, regulatory frameworks and policies must be established to define clear data protection and cybersecurity standards in the context of AI-6G integration. This includes setting guidelines for responsible AI usage, ensuring compliance with privacy regulations, e.g., general data protection regulation, and fostering international collaboration to combat cyber threats.

7.6 Human-AI interactive collaboration

The core challenges faced by human-AI interactive collaboration include trust, transparency, fairness, and ethics. The "black-box" nature of AI makes it difficult for humans to understand their decision-making logic, resulting in insufficient trust or overreliance. For instance, the complexity of DL models often makes it impossible for users to judge their reliability, thus affecting the effectiveness of collaboration [336]. Meanwhile, typically dynamic human-AI interactive collaboration requires high adaptability in task allocation and role switching. However, the existing collaboration frameworks are prone to unreasonable task allocation and delayed role switching [337]. Moreover, AI ethics and responsibility also need to be addressed urgently. For example, when an AI-driven decision is incorrect, it becomes crucial to establish a clear framework for assigning responsibility between humans and AI.

Through natural language explanations, visualization tools, and interactive interfaces, users can better understand the decision-making logic of AI and establish trust [338]. For example, the division of labor between humans and machines can be adjusted in real time according to task complexity and AI capabilities [339]. Coupled with an intuitive interface design, it can effectively reduce the difficulty for users to adapt to AI systems. For reducing bias and enhancing fairness, training with diverse and high-quality datasets and the application of fairness algorithms are crucial. Real-time bias-correction mechanisms can monitor and correct unfairness in AI-made decisions [340]. Furthermore, establishing a global ethical and legal framework will provide institutional guarantees for the development and application of AI. The boundaries of AI responsibility can be clarified through legislation and interdisciplinary ethical review mechanisms ensure the compliance of AI technology with social values and legal norms [341].

8 Future work

This section delves into directions that revolve around 6G, with the aim of providing robust support for its innovations. In-depth research on these areas will help lay a foundation for the development of 6G.

• **Adaptive learning mechanisms.** Adaptive learning mechanisms, leveraging AI, are set to become the cornerstone of smart network management. These mechanisms must handle the complexity and dynamism of 6G characterized by heterogeneous devices, variable traffic patterns, and the need for real-time processing. These models should be scalable to manage the exponential increase in connected devices and robust to ensure consistent performance under varying network states. One promising approach is DRL, which can dynamically optimize network decisions [272]. Integrating transfer learning techniques can enable AI models to apply knowledge gained in one domain to different yet relevant problems. Transfer learning can reduce the need for retraining models, saving computational resources and time. Another approach can be self-healing networks using AI. By incorporating predictive analytics, networks can anticipate failures and automatically reroute traffic or adjust parameters. This proactive approach to fault management will be critical for maintaining the high reliability and availability required by 6G applications.

• **Network architecture design.** With their promise of higher bandwidths and lower latencies, 6G networks provide an ideal platform for AI applications but pose challenges in architectural design. Future research can delve into creating network architectures that are inherently flexible and can dynamically adapt to the needs of AI applications. On the other hand, AI can significantly enhance network capability to manage these requirements by predicting traffic patterns and adjusting resources. A key component of such architectures is edge computing, which brings computational resources closer to the data sources, which is crucial for latency-sensitive AI applications. Research is expected to develop edge computing solutions that can seamlessly integrate with central cloud resources. Furthermore, deploying AI at the network edge introduces new challenges in distributed learning and data privacy. FL emerges as a viable solution, allowing AI models to be trained locally at the edge, with only model updates being shared to the central server.

• **Green AI and 6G networking.** The sustainability of 6G networks is critical environmentally and economically. AI has the potential to drive advancements in this area by optimizing network operations to reduce energy consumption. Future research can be devoted to AI models that can accurately predict the energy consumption of network components and optimize their operation to minimize energy use. This includes the dynamic adjustment of network infrastructure, such as BSs and data centers, based on user demand and traffic patterns. AI can enable these components to enter low-power states when

demand is low and quickly ramp up when the demand increases. Future research can also explore the design of new energy-efficient hardware that can support AI operations at the network edge. This includes the development of specialized processors that can run AI algorithms with high efficiency and low power consumption. Moreover, AI can manage the integration of renewable energy sources, e.g., solar and wind, and wireless power transfer into the network's power supply [342]. This involves predicting energy generation patterns and adjusting network operations to match the availability of renewable energy.

• **AI for ultra-low latency communications.** Ultra-low latency communication is a key aspect of 6G networks [343]. AI is poised to play a pivotal role in achieving this goal by providing real-time data processing and decision-making capabilities. Future research is expected to develop AI systems that can operate at the network edge, providing near-instantaneous response time for critical applications. This includes creating lightweight AI models that can be deployed on edge devices with limited computational resources. One challenge is the need for AI algorithms that can make decisions based on incomplete or uncertain data. The use of techniques, such as probabilistic modeling and approximate computing, is a promising direction to enable AI systems to operate effectively under these conditions. Another challenge is the development of AI-driven network orchestration tools that can allocate computational resources in real-time. These tools are expected to predict the computational needs of applications and allocate resources accordingly to ensure latency.

• **AI-enhanced cybersecurity in 6G.** AI has the potential to enhance the cybersecurity of 6G networks by providing advanced threat detection and response capabilities [344]. Future research can look into developing AI-based security frameworks that can identify and respond to cyber threats in real-time. This involves using ML algorithms to analyze network traffic and detect anomalies that may indicate a security breach. One challenge is the need for AI systems that can adapt to the constantly evolving landscape of cyber threats. Efforts can be directed to explore adaptive learning algorithms that can update their models in response to new types of attacks. Another challenge is developing AI-driven security protocols that can automatically patch vulnerabilities and respond to incidents without human intervention [345]. This requires the creation of intelligent systems that can understand the context of security incidents and take appropriate actions.

• **AI-driven 6G innovations.** The incorporation of AI within 6G networks can potentially drive technological innovations, including intelligent reflecting surface (IRS) [346], ISAC, O-RAN, non-terrestrial network (NTN), and near-field communication [347]. An IRS can dynamically augment wireless environments to optimize signal propagation, where AI can facilitate high-dimensional channel estimation and beamforming. ISAC merge sensing and communication functions to improve spectrum utilization and provide environmental context for network optimization, where AI can provide joint beamforming design and resource allocation for ISAC [348], and AI-based real-time data fusion can be essential to extract actionable insights from ISAC systems. ORAN can benefit from the AI-based network management of its software-defined architecture, enhancing adaptability and resilience. NTN, including satellite systems, can leverage AI for efficient resource management and fault detection. Interesting research directions can be AI models tailored to NTNs and addressing challenges like latency and dynamic network topology. By advancing AI applications in these novel areas, researchers can unlock new levels of performance and efficiency in 6G networks.

• **Ubiquitous computing.** With the proliferation of intelligent devices and the advancement of IoT technologies, the demand for intelligent services from users and devices continues to grow across wide areas. The integration of 6G and AI aims to provide AI services to users and devices worldwide, offering ubiquitous computing and meeting diverse intelligent service needs. By deploying edge servers on unmanned aerial vehicles (UAV) [349] or LEO satellites, multi-layered and heterogeneous computing services can be delivered, satisfying their differentiated QoS requirements. The primary challenge lies in designing rational, efficient, and dynamic resource allocation schemes for heterogeneous networks that integrate terrestrial BSs, UAVs, and LEO satellites. Compared to classical theories, such as convex optimization and Lyapunov optimization, particular emphasis needs to be placed on advanced algorithms like DRL to explore computation offloading strategies and resource allocation designs.

• **LLM-driven cognitive networking for 6G.** The rise of low-cost open-source LLMs, such as DeepSeek [350], is set to drastically lower barriers to AI-driven innovation. These accessible frameworks enable global researchers and organizations to rapidly prototype and deploy lightweight LLMs for 6G's mission-critical applications. The integration of LLMs with 6G networks is poised to revolutionize cognitive networking by enabling adaptive resource allocation [351], intent-based automation [352], embodied AI [353], and semantic communication [354]. Future research aims to embed LLMs into edge devices

through lightweight architectures for real-time network optimization, while addressing challenges such as computational overhead and privacy risks. Key directions include FL for distributed LLM training, cross-modal semantic encoding to reduce bandwidth consumption, and DT-aided network simulation. Innovations in sparsity-aware inference and hybrid quantum-classical frameworks may unlock scalable deployment across 6G infrastructures. However, balancing model compression with performance retention and ensuring robust decision-making under dynamic conditions remain critical hurdles.

9 Conclusion

In this all-encompassing review, we have initially highlighted that 6G network evolution towards an integrated wireless infrastructure platform amalgamating communication, sensing, computation, intelligence, and storage hinges on the profound integration of AI and communications. Concurrently, the exposition of the three-stage integration of AI within 6G networks, namely "AI4NET", "NET4AI", and "AIaaS", furnishes a lucid blueprint for understanding the technological development trajectory. These three consecutive stages present the strategic shift from utilizing AI to optimizing the network, through the network's facilitation of AI support to the network offering AIaaS, thereby underlining the interdependent and coordinated relationship between communication and AI. The 6G networks will introduce entirely new service paradigms, seamlessly integrating AI capabilities into various application scenarios, thereby serving as a powerful platform to support the advancement of AI. Wireless network large models are poised to serve as pivotal drivers in 6G systems, effectively integrating AI with communication technologies to advance intelligent, autonomous, and highly efficient network operations. By analyzing the standardization process of AI for wireless networks, we have highlighted the crucial milestones and ongoing efforts. We have further explored the challenges faced by the integration of AI and communications in the 6G era, including sustainability considerations, real-time AI demands, security and privacy issues, as well as human-AI interaction and collaboration. Finally, we have outlined promising future research opportunities that are expected to advance the development and optimization of AI and 6G communications. We are convinced that this review holds profound significance for the research on the integration of AI and communication for the 6G network, and it can provide deep insights for academic researchers and industry experts alike.

Acknowledgements This work was supported by Joint Funds for Regional Innovation and Development of National Natural Science Foundation of China (Grant No. U21A20449), Beijing Natural Science Foundation Program (Grant No. L232002), National Natural Science Foundation of China (Grant No. 62125108), National Key Research and Development Program of China (Grant No. 2020YFB1806804), and King Saud University, Riyadh, Saudi Arabia (Grant No. RSP2025R12).

References

1 Agiwal M, Roy A, Saxena N. Next generation 5G wireless networks: a comprehensive survey. IEEE Commun Surv Tut, 2016, 18: 1617–1655
2 Bhat J R, Alqahtani S A. 6G ecosystem: current status and future perspective. IEEE Access, 2021, 9: 43134–43167
3 Su Y, Liu Y, Zhou Y, et al. Broadband LEO satellite communications: architectures and key technologies. IEEE Wireless Commun, 2019, 26: 55–61
4 Yang H, Alphones A, Xiong Z, et al. Artificial-intelligence-enabled intelligent 6G networks. IEEE Netw, 2020, 34: 272–280
5 Behera R, Das K. A survey on machine learning: concept, algorithms and applications. Int J Innov Res Comput Commun Eng, 2017, 2: 2
6 Guo Y, Liu Y, Oerlemans A, et al. Deep learning for visual understanding: a review. Neurocomputing, 2016, 187: 27–48
7 Morgan D P, Scofield C L. Natural language processing. In: Neural Networks and Speech Processing. Berlin: Springer, 1991. 245–288
8 Wang C X, Renzo M D, Stanczak S, et al. Artificial intelligence enabled wireless networking for 5G and beyond: recent advances and future challenges. IEEE Wireless Commun, 2020, 27: 16–23
9 Li K, Lau B P L, Yuan X, et al. Toward ubiquitous semantic metaverse: challenges, approaches, and opportunities. IEEE Int Things J, 2023, 10: 21855–21872
10 Cai Q, Zhou Y, Liu L, et al. Query-aware semantic encoder-based resource allocation in task-oriented communications. IEEE Trans Mobile Comput, 2025. doi: 10.1109/TMC.2025.3541636
11 Zheng J, Li K, Mhaisen N, et al. Federated learning for online resource allocation in mobile edge computing: a deep reinforcement learning approach. In: Proceedings of IEEE Wireless Communications and Networking Conference (WCNC), Glasgow, 2023. 1–6
12 Peng Y, Tang X, Zhou Y, et al. Computing and communication cost-aware service migration enabled by transfer reinforcement learning for dynamic vehicular edge computing networks. IEEE Trans Mobile Comput, 2024, 23: 257–269

13 Arjunan T. Real-time detection of network traffic anomalies in big data environments using deep learning models. Int J Res Appl Sci Eng Tech, 2024, 12: 844–850
14 Cui Q, Ni W, Li S, et al. Learning-assisted clustered access of 5G/B5G networks to unlicensed spectrum. IEEE Wireless Commun, 2020, 27: 31–37
15 Guan X, Xu Z, Liu Y, et al. Reduction in energy consumption of the 5G communication system and beyond through collaborative optimization for BS site operation: challenges, efforts and the new approach. IEEE Trans Ind Inf, 2024, 20: 3948–3963
16 Zhang J, Cui Q, Zhang X, et al. Online optimization of energy-efficient user association and workload offloading for mobile edge computing. IEEE Trans Veh Technol, 2022, 71: 1974–1988
17 Cui Q, Zhang J, Zhang X, et al. Online anticipatory proactive network association in mobile edge computing for IoT. IEEE Trans Wireless Commun, 2020, 19: 4519–4534
18 Cui Q, Gong Z, Ni W, et al. Stochastic online learning for mobile edge computing: learning from changes. IEEE Commun Mag, 2019, 57: 63–69
19 Cai Q, Zhou Y, Liu L, et al. Collaboration of heterogeneous edge computing paradigms: how to fill the gap between theory and practice. IEEE Wireless Commun, 2024, 31: 110–117
20 Wang J, Jiang C, Kuang L. High-mobility satellite-UAV communications: challenges, solutions, and future research trends. IEEE Commun Mag, 2022, 60: 38–43
21 van den Broek F, Verdult R, De Ruiter J. Defeating IMSI catchers. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, New York, 2015. 340–351
22 Saleem R, Ni W, Ikram M, et al. Deep-reinforcement-learning-driven secrecy design for intelligent-reflecting-surface-based 6G-IoT networks. IEEE Int Things J, 2023, 10: 8812–8824
23 Dutta A, Hammad E. 5G security challenges and opportunities: a system approach. In: Proceedings of the 3rd 5G World Forum (5GWF), Bangalore, 2020. 109–114
24 Lyu X, Ren C, Ni W, et al. Online learning of optimal proactive schedule based on outdated knowledge for energy harvesting powered Internet-of-Things. IEEE Trans Wireless Commun, 2021, 20: 1248–1262
25 ITU. Framework and overall objectives of the future development of IMT for 2030 and beyond. Recommendation ITU-R M.2160-0, 2023. https://www.itu.int/rec/R-REC-M.2160-0-202311-I/en
26 Gong Z, Hashash O, Wang Y, et al. UAV-aided lifelong learning for AoI and energy optimization in nonstationary IoT networks. IEEE Int Things J, 2024, 11: 39206–39224
27 Li X, Cui Q, Xue Q, et al. A new batch access scheme with global QoS optimization for satellite-terrestrial networks. In: Proceedings of IEEE Global Communications Conference, Rio de Janeiro, 2022. 3929–3934
28 Cui Q, Zhang X, Ni W, et al. Big data analytics for intelligent management of autonomous vehicles in smart cities. In: Communication Technologies for Networked Smart Cities. London: IET, 2021
29 Cui Q, Wang Y, Chen K C, et al. Big data analytics and network calculus enabling intelligent management of autonomous vehicles in a smart city. IEEE Int Things J, 2019, 6: 2021–2034
30 Reinsel D, Gantz J, Rydning J. Data age 2025: the digitization of the world from edge to core. IDC White Paper, 2018. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf?gid=164649
31 Yue L, Chen T. AI large model and 6G network. In: Proceedings of IEEE Globecom Workshops (GC Wkshps), Kuala Lumpur, 2023. 2049–2054
32 Tong W, Li G Y. Nine challenges in artificial intelligence and wireless communications for 6G. IEEE Wireless Commun, 2022, 29: 140–145
33 Tao Z, Xu W, Huang Y, et al. Wireless network digital twin for 6G: generative AI as a key enabler. IEEE Wireless Commun, 2024, 31: 24–31
34 IMT-2030 (6G) Promotion Group. 6G network architecture outlook white paper (in Chinese). 2023. https://www.imt2030.org.cn/html/default/zhongwen/chengguofabu/baipishu/index.html?index=2
35 Yang T, Ning J, Lan D, et al. Kubeedge wireless for integrated communication and computing services everywhere. IEEE Wireless Commun, 2022, 29: 140–145
36 Chen X, Guo Z, Wang X, et al. Foundation model based native AI framework in 6G with cloud-edge-end collaboration. 2023. ArXiv:2310.17471
37 Wu J, Li R, An X, et al. Toward native artificial intelligence in 6G networks: system design, architectures, and paradigms. 2021. ArXiv:2103.02823
38 Liu G, Deng J, Zheng Q, et al. 6G native intelligence: technical challenges architecture and key features. Mobile Commun, 2021, 45: 68–78
39 Cao Y, Li S, Liu Y, et al. A comprehensive survey of AI-generated content (AIGC): a history of generative AI from GAN to ChatGPT. 2023. ArXiv:2303.04226
40 Maatouk A, Piovesan N, Ayed F, et al. Large language models for telecom: forthcoming impact on the industry. 2023. ArXiv:2308.06013
41 Jiang F, Peng Y, Dong L, et al. Large AI model-based semantic communications. IEEE Wireless Commun, 2024, 31: 68–75
42 DeepSeek-AI, Guo D, Yang D, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. 2025. ArXiv:2501.12948
43 6G Alliance of Network AI (6GANA). 6G network AI concept and terminology (in Chinese). 2022. https://www.6g-ana.com/upload/file/20220523/637889301757893555473508 1.pdf
44 Soury H, Smida B, Aliabadi S M. Accurate MMSE expressions for short-packet pilot-based channel estimation. IEEE Wireless Commun Lett, 2023, 12: 2188–2192
45 Farzamnia A, Hlaing N W, Haldar M K, et al. Channel estimation for sparse channel OFDM systems using least square and minimum mean square error techniques. In: Proceedings of International Conference on Engineering and Technology (ICET), Antalya, 2017. 1–5
46 Li J, Zhang Q, Xin X, et al. Deep learning-based massive MIMO CSI feedback. In: Proceedings of the 18th International Conference on Optical Communications and Networks (ICOCN), Huangshan, 2019. 1–3
47 Huawei Technologies Co., Ltd. R1-2304653: evaluation on AI/ML for CSI feedback enhancement. 2023. https://www.3gpp.org/ftp/tsg_ran/WG1_RL1/TSGR1_113/Docs
48 Guo J, Wen C K, Jin S, et al. Convolutional neural network-based multiple-rate compressive sensing for massive MIMO CSI feedback: design, simulation, and analysis. IEEE Trans Wireless Commun, 2020, 19: 2827–2840
49 Zhang Y, Zhang X, Liu Y. Deep learning based CSI compression and quantization with high compression ratios in FDD massive MIMO systems. IEEE Wireless Commun Lett, 2021, 10: 2101–2105
50 Wang Y, Sun J, Wang J, et al. Multi-rate compression for downlink CSI based on transfer learning in FDD massive MIMO systems. In: Proceedings of the 94th Vehicular Technology Conference (VTC2021-Fall), Norman, 2021. 1–5
51 Yu X, Li X, Wu H, et al. DS-NLCsiNet: exploiting non-local neural networks for massive MIMO CSI feedback. IEEE Commun Lett, 2020, 24: 2790–2794
52 Lu Z, Wang J, Song J. Multi-resolution CSI feedback with deep learning in massive MIMO system. In: Proceedings of IEEE International Conference on Communications (ICC), Dublin, 2020. 1–6

53 Gao X, Jin S, Wen C K, et al. ComNet: combination of deep learning and expert knowledge in OFDM receivers. IEEE Commun Lett, 2018, 22: 2627–2630
54 Pihlajasalo J, Korpi D, Honkala M, et al. Deep learning based OFDM physical-layer receiver for extreme mobility. In: Proceedings of the 55th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, 2021. 395–399
55 Mendonça M O K, Diniz P S R. OFDM receiver using deep learning: redundancy issues. In: Proceedings of the 28th European Signal Processing Conference (EUSIPCO), Amsterdam, 2021. 1687–1691
56 Chen W, Tang Z. Research on improved receiver of NOMA-OFDM signal based on deep learning. In: Proceedings of International Conference on Communications, Information System and Computer Engineering (CISCE), Beijing, 2021. 173–177
57 Balevi E, Andrews J G. One-bit OFDM receivers via deep learning. IEEE Trans Commun, 2019, 67: 4326–4336
58 Wang B, Xu K, Song P, et al. A deep learning-based intelligent receiver for OFDM. In: Proceedings of the 18th International Conference on Mobile Ad Hoc and Smart Systems (MASS), Denver, 2021. 562–563
59 An Y, Wu Z, Chunyang T. Performance of OFDM system receiver based on deep learning. In: Proceedings of International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), Xi'an, 2021. 571–574
60 Hussain M, Michelusi N. Throughput optimal beam alignment in millimeter wave networks. In: Proceedings of Information Theory and Applications Workshop (ITA), San Diego, 2017. 1–6
61 Zhu D, Choi J, Cheng Q, et al. High-resolution angle tracking for mobile wideband millimeter-wave systems with antenna array calibration. IEEE Trans Wireless Commun, 2018, 17: 7173–7189
62 Giordani M, Mezzavilla M, Barati C N, et al. Comparative analysis of initial access techniques in 5G mmWave cellular networks. In: Proceedings of Annual Conference on Information Science and Systems (CISS), Princeton, 2016. 268–273
63 Cui Q M, Zhang Z Y, Shi Y P, et al. Dynamic multichannel access based on deep reinforcement learning in distributed wireless networks. IEEE Syst J, 2022, 16: 5831–5834
64 Heng Y, Andrews J G. Machine learning-assisted beam alignment for mmWave systems. IEEE Trans Cogn Commun Netw, 2021, 7: 1142–1155
65 Yang J, Zhu W, Tao M, et al. Hierarchical beam alignment for millimeter-wave communication systems: a deep learning approach. IEEE Trans Wireless Commun, 2024, 23: 3541–3556
66 Heng Y, Mo J, Andrews J G. Learning site-specific probing beams for fast mmWave beam alignment. IEEE Trans Wireless Commun, 2022, 21: 5785–5800
67 Sayed A H, Yousef N R. Wireless Location. New York: Wiley, 2003
68 Luan N, Xiong K, Zhang Y, et al. 6G: typical applications, key technologies and challenges. Chin J Int Things, 2022, 6: 29–42
69 Maung N A M, Lwi B Y, Thida S. An enhanced RSS fingerprinting-based wireless indoor positioning using random forest classifier. In: Proceedings of International Conference on Advanced Information Technologies (ICAIT), Yangon, 2020. 59–63
70 Liao S, Zhao Q, Li J, et al. Outdoor positioning technology based on telecom data. Comput Eng Sci, 2018, 40: 649–659
71 Zhang L, Wang B, Xu Y, et al. Indoor location method based on RSSI probability distribution and CSI modified model. In: Proceedings of Global Conference on Robotics, Artificial Intelligence and Information Technology (GCRAIT), Chicago, 2022. 429–433
72 Wang Q, Zou S, Sun Y, et al. Toward intelligent and adaptive task scheduling for 6G: an intent-driven framework. IEEE Trans Cogn Commun Netw, 2024, 10: 1975–1988
73 Saad W, Bennis M, Chen M. A vision of 6G wireless systems: applications, trends, technologies, and open research problems. IEEE Netw, 2020, 34: 134–142
74 Wang J, Tang J, Xu Z, et al. Spatiotemporal modeling and prediction in cellular networks: a big data enabled deep learning approach. In: Proceedings of IEEE Conference on Computer Communications, Atlanta, 2017. 1–9
75 Nie L, Jiang D, Yu S, et al. Network traffic prediction based on deep belief network in wireless mesh backbone networks. In: Proceedings of IEEE Wireless Communications and Networking Conference (WCNC), San Francisco, 2017. 1–5
76 Jaffry S. Cellular traffic prediction with recurrent neural network. 2020. ArXiv:2003.02807
77 Zhang C, Dang S, Shihada B, et al. Dual attention-based federated learning for wireless traffic prediction. In: Proceedings of IEEE Conference on Computer Communications, Vancouver, 2021. 1–10
78 Zhang L, Zhang C, Shihada B. Efficient wireless traffic prediction at the edge: a federated meta-learning approach. IEEE Commun Lett, 2022, 26: 1573–1577
79 Zhang C, Zhang H, Qiao J, et al. Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data. IEEE J Sel Areas Commun, 2019, 37: 1389–1401
80 Zhao Y, Li M, Lai L, et al. Federated learning with non-IID data. 2018. ArXiv:1806.00582
81 Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of International Conference on Machine Learning, Sydney, 2017. 1126–1135
82 Inamdar M A, Kumaraswamy H V. Energy efficient 5G networks: techniques and challenges. In: Proceedings of International Conference on Smart Electronics and Communication (ICOSEC), Trichy, 2020. 1317–1322
83 Peng J, Hong P, Xue K. Stochastic analysis of optimal base station energy saving in cellular networks with sleep mode. IEEE Commun Lett, 2014, 18: 612–615
84 Debaillie B, Desset C, Louagie F. A flexible and future-proof power model for cellular base stations. In: Proceedings of IEEE Vehicular Technology Conference (VTC Spring), Glasgow, 2015. 1–7
85 Salem F E, Altman Z, Gati A, et al. Reinforcement learning approach for advanced sleep modes management in 5G networks. In: Proceedings of IEEE Vehicular Technology Conference (VTC-Fall), Chicago, 2018. 1–5
86 Masoudi M, Khafagy M G, Soroush E, et al. Reinforcement learning for traffic-adaptive sleep mode management in 5G networks. In: Proceedings of the 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications, London, 2020. 1–6
87 Lin S, Qiu C, Tan J, et al. DADEs: 5G dual-adaptive delay-aware and energy-saving system with tandem learning. In: Proceedings of IEEE Global Communications Conference, Rio de Janeiro, 2022. 1–6
88 Li T, Yu L, Ma Y, et al. Carbon emissions of 5G mobile networks in China. Nat Sustain, 2023, 6: 1620–1631
89 McHenry M, Livsics E, Nguyen T, et al. XG dynamic spectrum access field test results. IEEE Commun Mag, 2007, 45: 51–57
90 Kaur A, Kumar K. A comprehensive survey on machine learning approaches for dynamic spectrum access in cognitive radio networks. J Exp Theor Artif Intell, 2022, 34: 1–40
91 Alabi C A, Imoize A L, Giwa M A, et al. Artificial intelligence in spectrum management: policy and regulatory considerations. In: Proceedings of International Conference on Multidisciplinary Engineering and Applied Science (ICMEAS), Abuja, 2023. 1–6
92 Liu Y, Bi S, Shi Z, et al. When machine learning meets big data: a wireless communication perspective. IEEE Veh Technol Mag, 2020, 15: 63–72
93 Chernogorov F, Puttonen J. User satisfaction classification for minimization of drive tests QoS verification. In: Proceedings of the 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), London, 2013. 2165–2169

94 Luo Z Q, Zheng X, López-Pérez D, et al. SRCON: a data-driven network performance simulator for real-world wireless networks. IEEE Commun Mag, 2023, 61: 96–102
95 Mao Q, Hu F, Hao Q. Deep learning for intelligent wireless networks: a comprehensive survey. IEEE Commun Surv Tut, 2018, 20: 2595–2621
96 You X H, Pan Z W, Cao X Q, et al. The 5G mobile communication: the development trends and its emerging key techniques (in Chinese). Sci Sin Inform, 2014, 44: 551–563
97 Tu Y H, Ma Y W, Li Z X, et al. Applying deep reinforcement learning for self-organizing network architecture. In: Proceedings of the 6th International Conference on Knowledge Innovation and Invention (ICKII), Sapporo, 2023. 16–20
98 Wang W, Duan X, Sun W, et al. Research on mobility prediction in 5G and beyond for vertical industries. In: Proceedings of IEEE/CIC International Conference on Communications in China (ICCC Workshops), Xiamen, 2021. 379–383
99 Shubyn B, Lutsiv N, Syrotynskyi O, et al. Deep learning based adaptive handover optimization for ultra-dense 5G mobile networks. In: Proceedings of International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), Lviv-Slavske, 2020. 869–872
100 Liu Q, Chuai G, Wang J, et al. Proactive mobility management with trajectory prediction based on virtual cells in ultra-dense networks. IEEE Trans Veh Technol, 2020, 69: 8832–8842
101 Prado A, Stöckeler F, Mehmeti F, et al. Enabling proportionally-fair mobility management with reinforcement learning in 5G networks. IEEE J Sel Areas Commun, 2023, 41: 1845–1858
102 Yan L, Ding H, Zhang L, et al. Machine learning-based handovers for sub-6 GHz and mmWave integrated vehicular networks. IEEE Trans Wireless Commun, 2019, 18: 4873–4885
103 Abdellatif A A, Abo-Eleneen A, Mohamed A, et al. Intelligent-slicing: an AI-assisted network slicing framework for 5G-and-beyond networks. IEEE Trans Netw Serv Manage, 2023, 20: 1024–1039
104 Qin X, Li Y, Song X, et al. Timeliness of information for computation-intensive status updates in task-oriented communications. IEEE J Sel Areas Commun, 2023, 41: 623–638
105 Wang Y, Chen K C, Gong Z, et al. Reliability-guaranteed uplink resource management in proactive mobile network for minimal latency communications. IEEE Trans Wireless Commun, 2023, 22: 5018–5030
106 Zhang P, Su Y, Wang J, et al. Reinforcement learning assisted bandwidth aware virtual network resource allocation. IEEE Trans Netw Serv Manage, 2022, 19: 4111–4123
107 Zhang P, Chen N, Xu G, et al. Multi-target-aware dynamic resource scheduling for cloud-fog-edge multi-tier computing network. IEEE Trans Intell Transp Syst, 2024, 25: 3885–3897
108 Nguyen V D, Vu T X, Nguyen N T, et al. Network-aided intelligent traffic steering in 6G O-RAN: a multi-layer optimization framework. IEEE J Sel Areas Commun, 2024, 42: 389–405
109 Sami H, Otrok H, Bentahar J, et al. AI-based resource provisioning of IoE services in 6G: a deep reinforcement learning approach. IEEE Trans Netw Serv Manage, 2021, 18: 3527–3540
110 Dong T, Zhuang Z, Qi Q, et al. Intelligent joint network slicing and routing via GCN-powered multi-task deep reinforcement learning. IEEE Trans Cogn Commun Netw, 2022, 8: 1269–1286
111 Mei J, Wang X, Zheng K, et al. Intelligent radio access network slicing for service provisioning in 6G: a hierarchical deep reinforcement learning approach. IEEE Trans Commun, 2021, 69: 6063–6078
112 Bhattacharya P, Patel F, Alabdulatif A, et al. A deep-Q learning scheme for secure spectrum allocation and resource management in 6G environment. IEEE Trans Netw Serv Manage, 2022, 19: 4989–5005
113 Fan J, Mu D, Liu Y. Research on network traffic prediction model based on neural network. In: Proceedings of the 2nd International Conference on Information Systems and Computer Aided Education (ICISCAE), Dalian, 2019. 554–557
114 Wu J, Ota K, Dong M, et al. Big data analysis-based security situational awareness for smart grid. IEEE Trans Big Data, 2018, 4: 408–417
115 Yin K, Yang Y, Yao C, et al. Long-term prediction of network security situation through the use of the transformer-based model. IEEE Access, 2022, 10: 56145–56157
116 Xu Z, Liu J, Luo X, et al. Cross-version defect prediction via hybrid active learning with kernel principal component analysis. In: Proceedings of the 25th International Conference on Software Analysis, Evolution and Reengineering (SANER), Campobasso, 2018. 209–220
117 Zhang Z, Li J. Application of active learning strategies in fault detection and diagnosis within communication networks. In: Proceedings of the 7th World Conference on Computing and Communication Technologies (WCCCT), Chengdu, 2024. 186–193
118 Iswarya P, Manikandan K. Algorithms for fault detection and diagnosis in wireless sensor networks using deep learning and machine learning — an overview. In: Proceedings of the 10th International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, 2024. 1404–1409
119 Liu G Y, Zhang H M, Tong Z, et al. 6G mobile information network architecture: migrate from communication to XaaS (in Chinese). Sci Sin Inform, 2024, 54: 1236–1266
120 China Telecom Research Institute. 6G vision and technology white paper (in Chinese). 2022. http://doc.cserver.com.cn/doc_d5e51a1e-3c89-4c9d-ae4a-34e39f569cd3.html
121 China Unicom Research Institute. China unicom 6G white paper (in Chinese). 2021. http://221.179.172.81/images/20210322/30691616408868127.pdf
122 Wu J, Deng J, Peng C, et al. Task-centered 6G network AI architecture. Radio Commun Tech, 2022, 48: 599–613
123 CICT, DTmobile. Full coverage, scene intelligence — 6G vision and technology trends white paper (in Chinese). 2020. https://www.5gxt.com/material/data1812.htm
124 CICT, CICT Mobile. 6G network architecture white paper (in Chinese). 2022. https://www.cictmobile.com/upload/File/202302/20194e9d54af46b79a4e06c80773f620.pdf
125 OPPO Research Institute. 6G AI-Cube intelligent network (in Chinese). 2021. https://www.oppo.com/content/dam/oppo/cn/mkt/newsroom/press/455/whitepaper.pdf
126 Ericsson. Co-creating a cyber-physical world. 2022. https://www.ericsson.com/en/reports-and-papers/white-papers/a-research-outlook-towards-6g
127 Hoydis J, Aoudia F A, Valcarce A, et al. Toward a 6G AI-native air interface. IEEE Commun Mag, 2021, 59: 76–81
128 Consortium B G P. Beyond 5G white paper — version 1.0. 2022. https://b5g.jp/w/wp-content/uploads/pdf/whitepaper_en_1-0.pdf
129 Zhang P, Niu K, Tian H, et al. Technology prospect of 6G mobile communications. J Commun, 2019, 40: 141–148
130 Beijing University of Posts and Telecommunications. 6G green wireless access network white paper for multi-dimensional stereo and full scenarios (in Chinese). 2023. https://www.fxbaogao.com/detail/3944018
131 Southeast University, Purple Mountain Laboratories. 6G research white paper (in Chinese). 2020. https://ncrl.seu.edu.cn/_upload/article/files/17/86/577c0170409eae5399c8fbdf835f/e64c829e-5db9-4d2a-93e5-6bb60d145f0d.pdf
132 Zhang L, Liang Y C, Niyato D. 6G visions: mobile ultra-broadband, super Internet-of-Things, and artificial intelligence. China Commun, 2019, 16: 1–14
133 Luo H, Zhang T, Zhao C, et al. Integrated sensing and communications framework for 6G networks. 2024. ArXiv:2405.19925
134 Tao M, Zhou Y, Shi Y, et al. Federated edge learning for 6G: foundations, methodologies, and applications. Proc IEEE,

2024. doi: 10.1109/JPROC.2024.3509739
135 Yang Y, Ma M, Wu H, et al. 6G network AI architecture for everyone-centric customized services. IEEE Netw, 2022, 37: 71–80
136 Lu Y, Maharjan S, Zhang Y. Adaptive edge association for wireless digital twin networks in 6G. IEEE Int Things J, 2021, 8: 16219–16230
137 Taleb T, Aguiar R L, Yahia I G B, et al. White paper on 6G networking. 2020. http://urn.fi/urn:isbn:9789526226842
138 Liu G, Deng J, Zheng Q, et al. 6G native intelligence: technical challenges, architecture and key features. Mobile Commun, 2021, 45: 68–78
139 Liu G, Deng J, Li N, et al. Native AI and service based architecture for 6G wireless network. Radio Commun Tech, 2022, 48: 562–573
140 3rd Generation Partership Project (3GPP). Service requirements for the 5G system. TS 22.261. https://www.3gpp.org/ftp/Specs/archive/22_series/22.261
141 3rd Generation Partership Project (3GPP). Study on traffic characteristics and performance requirements for AI/ML model transfer in 5GS. TR 22.874. https://www.3gpp.org/ftp/Specs/archive/22_series/22.874
142 6G Alliance of Network AI (6GANA). 6G data service — concept and requirements (in Chinese). 2022. https://www.6g-ana.com/upload/file/20220523/637889302817268731197 1492.pdf
143 Alkhateeb A. DeepMIMO: a generic deep learning dataset for millimeter wave and massive MIMO applications. 2019. ArXiv:1902.06435
144 O'Shea T J, Corgan J, Clancy T C. Convolutional radio modulation recognition networks. 2016. ArXiv:1602.04105
145 ITU. AI/ML in 5G challenge. 2023. https://www.itu.int/en/ITU-T/AI/challenge/Pages/default.aspx
146 Lyu X, Ren C, Ni W, et al. Distributed online learning of cooperative caching in edge cloud. IEEE Trans Mobile Comput, 2021, 20: 2550–2562
147 Liu Y, Zhang W, Li L, et al. Toward autonomous trusted networks-from digital twin perspective. IEEE Netw, 2024, 38: 84–91
148 NGMN. 6G requirements and design considerations (white paper). 2023. https://www.ngmn.org/wp-content/uploads/NGMN_6G_Requirements_and_Design_Considerations.pdf
149 Cui Q, Zhu Z, Ni W, et al. Edge-intelligence-empowered, unified authentication and trust evaluation for heterogeneous beyond 5G systems. IEEE Wireless Commun, 2021, 28: 78–85
150 Li K, Cui Q, Zhu Z, et al. Lightweight, privacy-preserving handover authentication for integrated terrestrial-satellite networks. In: Proceedings of IEEE International Conference on Communications, Seoul, 2022. 25–31
151 Hu S, Chen X, Ni W, et al. Distributed machine learning for wireless communication networks: techniques, architectures, and applications. IEEE Commun Surv Tut, 2021, 23: 1458–1493
152 6G Alliance of Network AI (6GANA). White paper on native AI technology requirements of 6G network (in Chinese). 2022. https://www.6g-ana.com/upload/file/20220523/637889301773049743 4706068.pdf
153 Yuan X, Ni W, Ding M, et al. Amplitude-varying perturbation for balancing privacy and utility in federated learning. IEEE Trans Inform Forensic Secur, 2023, 18: 1884–1897
154 Zhao X, Cui Q, Li W, et al. Convergence-privacy-fairness trade-off in personalized federated learning. Trans Mach Learn Comm Netw, 2025, 3: 246–262
155 Xiao B, Yu X, Ni W, et al. Over-the-air federated learning: status quo, open challenges, and future directions. 2024. ArXiv:2307.00974
156 Yu X, Xiao B, Ni W, et al. Optimal adaptive power control for over-the-air federated edge learning under fading channels. IEEE Trans Commun, 2023, 71: 5199–5213
157 Sun P, Liu E, Ni W, et al. Reconfigurable intelligent surface-assisted wireless federated learning with imperfect aggregation. IEEE Trans Commun, 2025, 73: 1058–1071
158 Li W, Lv T, Ni W, et al. Decentralized federated learning over imperfect communication channels. IEEE Trans Commun, 2024, 72: 6973–6991
159 Li X, Huang K, Yang W, et al. On the convergence of FedAvg on non-IID data. 2019. ArXiv:1907.02189
160 Luo B, Xiao W, Wang S, et al. Tackling system and statistical heterogeneity for federated learning with adaptive client sampling. In: Proceedings of IEEE Conference on Computer Communications, London, 2022. 1739–1748
161 Wang S, Perazzone J, Ji M, et al. Federated learning with flexible control. In: Proceedings of IEEE Conference on Computer Communications, New York, 2023. 1–10
162 Zhao B, Cui Q, Liang S, et al. Green concerns in federated learning over 6G. China Commun, 2022, 19: 50–69
163 Liang S, Cui Q, Huang X, et al. Efficient hierarchical federated services for heterogeneous mobile edge. IEEE Trans Serv Comput, 2025, 18: 140–155
164 Fallah A, Mokhtari A, Ozdaglar A. Personalized federated learning with theoretical guarantees: a model-agnostic meta-learning approach. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020. 3557–3568
165 Tan A Z, Yu H, Cui L, et al. Towards personalized federated learning. IEEE Trans Neural Netw Learn Syst, 2023, 34: 9587–9603
166 Passerat-Palmbach J, Farnan T, McCoy M, et al. Blockchain-orchestrated machine learning for privacy preserving federated learning in electronic health data. In: Proceedings of IEEE International Conference on Blockchain, Rhodes, 2020. 550–555
167 Połap D, Srivastava G, Yu K. Agent architecture of an intelligent medical system based on federated learning and blockchain technology. J Inf Secur Appl, 2021, 58: 102748
168 Wang K I K, Zhou X, Liang W, et al. Federated transfer learning based cross-domain prediction for smart manufacturing. IEEE Trans Ind Inf, 2022, 18: 4088–4096
169 Wan Y, Qu Y, Ni W, et al. Data and model poisoning backdoor attacks on wireless federated learning, and the defense mechanisms: a comprehensive survey. IEEE Commun Surv Tut, 2024, 26: 1861–1897
170 Li K, Zheng J, Yuan X, et al. Data-agnostic model poisoning against federated learning: a graph autoencoder approach. IEEE Trans Inform Forensic Secur, 2024, 19: 3465–3480
171 Liu Y, Kang Y, Zou T, et al. Vertical federated learning: concepts, advances, and challenges. IEEE Trans Knowl Data Eng, 2024, 36: 3615–3634
172 Zhang J, Jiang Y. A vertical federation recommendation method based on clustering and latent factor model. In: Proceedings of International Conference on Electronic Information Engineering and Computer Science (EIECS), Changchun, 2021. 362–366
173 Tang F, Liang S, Ling G, et al. IHVFL: a privacy-enhanced intention-hiding vertical federated learning framework for medical data. Cybersecurity, 2023, 6: 37
174 Yuan X, Chen J, Yang J, et al. FedSTN: graph representation driven federated learning for edge computing enabled urban traffic flow prediction. IEEE Trans Intell Transp Syst, 2022, 24: 8738–8748
175 Chu T, Wang J, Codeca L, et al. Multi-agent deep reinforcement learning for large-scale traffic signal control. IEEE Trans Intell Transp Syst, 2019, 21: 1086–1095
176 Wu T, Zhou P, Liu K, et al. Multi-agent deep reinforcement learning for urban traffic light control in vehicular networks.

IEEE Trans Veh Technol, 2020, 69: 8243–8256
177 Palanisamy P. Multi-agent connected autonomous driving using deep reinforcement learning. In: Proceedings of International Joint Conference on Neural Networks (IJCNN), Glasgow, 2020. 1–7
178 Yu K, Cui Q, Lyu X, et al. Efficient collaborative computing for multilayer LEO satellites with spatiotemporal dynamics: a long-term continuous timescale optimization. IEEE Int Things J, 2025, 12: 7459–7471
179 Zhou T, Tang D, Zhu H, et al. Multi-agent reinforcement learning for online scheduling in smart factories. Robot Comput-Integr Manuf, 2021, 72: 102202
180 Cao Z, Zhou P, Li R, et al. Multiagent deep reinforcement learning for joint multichannel access and task offloading of mobile-edge computing in Industry 4.0. IEEE Int Things J, 2020, 7: 6201–6213
181 Duan Q, Hu S, Deng R, et al. Combined federated and split learning in edge computing for ubiquitous intelligence in Internet of Things: state-of-the-art and future directions. Sensors, 2022, 22: 5983
182 Lyu X, Li Y, He Y, et al. Objective-driven differentiable optimization of traffic prediction and resource allocation for split AI inference edge networks. Trans Mach Learn Comm Netw, 2024, 2: 1178–1192
183 Lin Z, Qu G, Chen X, et al. Split learning in 6G edge networks. IEEE Wireless Commun, 2024, 31: 170–176
184 Vepakomma P, Gupta O, Swedish T, et al. Split learning for health: distributed deep learning without sharing raw patient data. 2018. ArXiv:1812.00564
185 Poirot M G, Vepakomma P, Chang K, et al. Split learning for collaborative deep learning in healthcare. 2019. ArXiv:1912.12115
186 Li Z, Yan C, Zhang X, et al. Split learning for distributed collaborative training of deep learning models in health informatics. In: Proceedings of AMIA Annual Symposium Proceedings, 2024. 1047–1056
187 Jeon J, Kim J. Privacy-sensitive parallel split learning. In: Proceedings of International Conference on Information Networking (ICOIN), Barcelona, 2020. 7–9
188 Wu W, Li M, Qu K, et al. Split learning over wireless networks: parallel design and resource management. IEEE J Sel Areas Commun, 2023, 41: 1051–1066
189 Mao Y, Yu X, Huang K, et al. Green edge AI: a contemporary survey. Proc IEEE, 2024, 112: 880–911
190 Nguyen H X, Trestian R, To D, et al. Digital twin for 5G and beyond. IEEE Commun Mag, 2021, 59: 10–15
191 Grieves M W. Product lifecycle management: the new paradigm for enterprises. IJPD, 2005, 2: 71–84
192 Cao Y, Dai L, Tan J, et al. Advancing ubiquitous wireless connectivity through channel twinning. IEEE Commun Mag, 2025. doi: 10.1109/MCOM.001.2400476
193 Tao F, Zhang H, Liu A, et al. Digital twin in industry: state-of-the-art. IEEE Trans Ind Inf, 2018, 15: 2405–2415
194 Cui Y, Lv T, Ni W, et al. Digital twin-aided learning for managing reconfigurable intelligent surface-assisted, uplink, user-centric cell-free systems. IEEE J Sel Areas Commun, 2023, 41: 3175–3190
195 Barricelli B R, Casiraghi E, Fogli D. A survey on digital twin: definitions, characteristics, applications, and design implications. IEEE Access, 2019, 7: 167653
196 Li Z, Duan M, Xiao B, et al. A novel anomaly detection method for digital twin data using deconvolution operation with attention mechanism. IEEE Trans Ind Inf, 2022, 19: 7278–7286
197 Qin B, Pan H, Dai Y, et al. Machine and deep learning for digital twin networks: a survey. IEEE Int Things J, 2024, 11: 34694–34716
198 Wu X, Lian W, Zhou M, et al. A digital twin-based fault diagnosis framework for bogies of high-speed trains. IEEE J Radio Freq Identif, 2022, 7: 203–207
199 Cui C, Ma Y, Cao X, et al. Human-autonomy teaming on autonomous vehicles with large language model-enabled human digital twins. In: Proceedings of IEEE/ACM Symposium on Edge Computing (SEC), Wilmington, 2023. 319–324
200 Liu T, Tang L, Wang W, et al. Digital-twin-assisted task offloading based on edge collaboration in the digital twin edge network. IEEE Int Things J, 2021, 9: 1427–1444
201 Huang Z, Li D, Cai J, et al. Collective reinforcement learning based resource allocation for digital twin service in 6G networks. J Netw Comput Appl, 2023, 217: 103697
202 Wang J, Zhang J, Zhang Y, et al. Towards 6G digital twin channel using radio environment knowledge pool. 2023. ArXiv:2312.10287
203 Wang H, Zhang J, Nie G, et al. Digital twin channel for 6G: concepts, architectures and potential applications. 2024. ArXiv:2403.12467
204 Nie G, Zhang J, Zhang Y, et al. A predictive 6G network with environment sensing enhancement: from radio wave propagation perspective. China Commun, 2022, 19: 105–122
205 Miao Y, Zhang Y, Zhang J, et al. Demo abstract: predictive radio environment for digital twin communication platform via enhanced sensing. In: Proceedings of IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), New York, 2023. 1–2
206 Zhang J. The interdisciplinary research of big data and wireless channel: a cluster-nuclei based channel model. China Commun, 2016, 13: 14–26
207 Wang J, Zhang J, Sun Y, et al. Electromagnetic wave property inspired radio environment knowledge construction and AI-based verification for 6G digital twin channel. 2024. ArXiv:2406.00690
208 ITU. Digital twin network — requirements and architecture. 2022. https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-Y.3090-202202-I!!PDF-E&type=items
209 Jiang Y, Ma B, Wang X, et al. Blockchained federated learning for Internet of Things: a comprehensive survey. ACM Comput Surv, 2024, 56: 1–37
210 van Huynh D, Nguyen V D, Khosravirad S R, et al. URLLC edge networks with joint optimal user association, task offloading and resource allocation: a digital twin approach. IEEE Trans Commun, 2022, 70: 7669–7682
211 Khan L U, Han Z, Saad W, et al. Digital twin of wireless systems: overview, taxonomy, challenges, and opportunities. IEEE Commun Surv Tut, 2022, 24: 2230–2254
212 Lu Y, Huang X, Zhang K, et al. Communication-efficient federated learning and permissioned blockchain for digital twin edge networks. IEEE Int Things J, 2020, 8: 2276–2288
213 Zhang K, Cao J, Zhang Y. Adaptive digital twin and multiagent deep reinforcement learning for vehicular edge computing and networks. IEEE Trans Ind Inf, 2021, 18: 1405–1413
214 Hu C, Fan W, Zeng E, et al. Digital twin-assisted real-time traffic data prediction method for 5G-enabled Internet of Vehicles. IEEE Trans Ind Inf, 2021, 18: 2811–2819
215 Chen D, Lv Z. Artificial intelligence enabled digital twins for training autonomous cars. Int Things Cyber-Phys Syst, 2022, 2: 31–41
216 Liu K, Xu X, Dai P, et al. Cooperative sensing and uploading for quality-cost tradeoff of digital twins in VEC. IEEE Trans Consumer Electron, 2024, 70: 3614–3625
217 IMT-2030 (6G) Promotion Group. 6G network architecture vision and key technology outlook white paper (in Chinese). 2021. https://www.imt2030.org.cn/
218 Yukun S, Bo L, Junlin L, et al. Computing power network: a survey. China Commun, 2024, 21: 109–145
219 Jia Q M, Guo K, Zhou X M, et al. Design and discussion for new computing power network architecture. Inform Commun

Technol Policy, 2022, 48: 18–23
220 Yao H, Geng L. Trend of next generation network architecture: computing and networking convergence evolution. Telecommun Sci, 2019, 35: 38–43
221 Tang X, Cao C, Wang Y, et al. Computing power network: the architecture of convergence of computing and networking towards 6G requirement. China Commun, 2021, 18: 175–185
222 Xiao D, Zhang J A, Liu X, et al. A two-stage GCN-based deep reinforcement learning framework for SFC embedding in multi-datacenter networks. IEEE Trans Netw Serv Manage, 2023, 20: 4297–4312
223 Wang X, Ren X, Qiu C, et al. Net-in-AI: a computing-power networking framework with adaptability, flexibility, and profitability for ubiquitous AI. IEEE Netw, 2020, 35: 280–288
224 Tokusashi Y, Dang H T, Pedone F, et al. The case for in-network computing on demand. In: Proceedings of the 14th EuroSys Conference, Dresden, 2019. 1–16
225 Ren X, Qiu C, Wang X, et al. AI-Bazaar: a cloud-edge computing power trading framework for ubiquitous AI services. IEEE Trans Cloud Comput, 2022, 11: 2337–2348
226 Duan L, Sun Y, Ni W, et al. Attacks against cross-chain systems and defense approaches: a contemporary survey. IEEE CAA J Autom Sin, 2023, 10: 1647–1667
227 Duan L, Yang L, Liu C, et al. A new smart contract anomaly detection method by fusing opcode and source code features for blockchain services. IEEE Trans Netw Serv Manage, 2023, 20: 4354–4368
228 Liu J, Sun Y, Su J, et al. Computing power network: a testbed and applications with edge intelligence. In: Proceedings of IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), New York, 2022. 1–2
229 Li M X, Cao C, Tang X Y, et al. Research on edge resource scheduling solutions for computing power network (in Chinese). Front Data Comput, 2020, 2: 80–91
230 Chen Q, Yang C, Lan S, et al. Two-stage evolutionary search for efficient task offloading in edge computing power networks. IEEE Int Things J, 2024, 11: 30787–30799
231 Sodhro A H, Sodhro G H, Guizani M, et al. AI-enabled reliable channel modeling architecture for fog computing vehicular networks. IEEE Wireless Commun, 2020, 27: 14–21
232 Deng C, Fang X, Wang X. UAV-enabled mobile-edge computing for AI applications: joint model decision, resource allocation, and trajectory optimization. IEEE Int Things J, 2022, 10: 5662–5675
233 Yao A C. Protocols for secure computations. In: Proceedings of the 23rd Annual Symposium on Foundations of Computer Science (SFCS), Chicago, 1982. 160–164
234 Okamura T, Teranishi I. Enhancing fintech security with secure multi-party computation technology. NEC Tech J, 2017, 11: 46–50
235 Shamir A. How to share a secret. Commun ACM, 1979, 22: 612–613
236 Haque A, Heath D, Kolesnikov V, et al. Garbled circuits with sublinear evaluator. In: Proceedings of the 41st Annual International Conference on the Theory and Application of Cryptographic Technology, 2022. 37–64
237 Lindell Y, Pinkas B. A proof of security of Yao's protocol for two-party computation. J Cryptol, 2009, 22: 161–188
238 Acar A, Aksu H, Uluagac A S, et al. A survey on homomorphic encryption schemes. ACM Comput Surv, 2018, 51: 1–35
239 Wang R, Lai J, Zhang Z, et al. Privacy-preserving federated learning for Internet of Medical Things under edge computing. IEEE J Biomed Health Inform, 2022, 27: 854–865
240 Song J, Wang W, Gadekallu T R, et al. EPPDA: an efficient privacy-preserving data aggregation federated learning scheme. IEEE Trans Netw Sci Eng, 2022, 10: 3047–3057
241 Zhu H, Mong Goh R S, Ng W K. Privacy-preserving weighted federated learning within the secret sharing framework. IEEE Access, 2020, 8: 198275
242 Huang A, Liu Y, Chen T, et al. StarFL: hybrid federated learning architecture for smart urban computing. ACM Trans Intell Syst Technol, 2021, 12: 1–23
243 Tang X, Zhu L, Shen M, et al. Secure and trusted collaborative learning based on blockchain for artificial intelligence of things. IEEE Wireless Commun, 2022, 29: 14–22
244 Yu X, Lv T, Li W, et al. Multi-task semantic communication with graph attention-based feature correlation extraction. IEEE Trans Mobile Comput, 2025, : 1–18
245 Niu K, Zhang P. A mathematical theory of semantic communication. 2024. ArXiv:2401.13387
246 Zhang P, Xu X, Dong C, et al. Model division multiple access for semantic communications. Front Inform Technol Electron Eng, 2023, 24: 801–812
247 Yang W, Du H, Liew Z Q, et al. Semantic communications for future Internet: fundamentals, applications, and challenges. IEEE Commun Surv Tut, 2022, 25: 213–250
248 Xie H, Qin Z, Li G Y, et al. Deep learning enabled semantic communication systems. IEEE Trans Signal Process, 2021, 69: 2663–2675
249 Lu K, Li R, Chen X, et al. Reinforcement learning-powered semantic communication via semantic similarity. 2021. ArXiv:2108.12121
250 Jiang P, Wen C K, Jin S, et al. Deep source-channel coding for sentence semantic transmission with HARQ. IEEE Trans Commun, 2022, 70: 5225–5240
251 Weng Z, Qin Z. Semantic communication systems for speech transmission. IEEE J Sel Areas Commun, 2021, 39: 2434–2444
252 Weng Z, Qin Z, Li G Y. Semantic communications for speech signals. In: Proceedings of IEEE International Conference on Communications, Montreal, 2021. 1–6
253 Tong H, Yang Z, Wang S, et al. Federated learning for audio semantic communication. Front Comms Net, 2021, 2: 734402
254 Bourtsoulatze E, Kurka D B, Gunduz D. Deep joint source-channel coding for wireless image transmission. IEEE Trans Cogn Commun Netw, 2019, 5: 567–579
255 Wang S, Dai J, Liang Z, et al. Wireless deep video semantic transmission. IEEE J Sel Areas Commun, 2022, 41: 214–229
256 Yenduri G, Ramalingam M, Selvi G C, et al. GPT (generative pre-trained transformer) — a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. IEEE Access, 2024, 12: 54608–54649
257 Chowdhury M Z, Shahjalal M, Ahmed S, et al. 6G wireless communication systems: applications, requirements, technologies, challenges, and research directions. IEEE Open J Commun Soc, 2020, 1: 957–975
258 Tong W, Peng C, Yang T, et al. Ten issues of NetGPT. 2023. ArXiv:2311.13106
259 Shao J, Tong J, Wu Q, et al. WirelessLLM: empowering large language models towards wireless intelligence. J Commun Inf Netw, 2024, 9: 99–112
260 Zou H, Zhao Q, Tian Y, et al. TelecomGPT: a framework to build telecom-specfic large language models. 2024. ArXiv:2407.09424
261 Letaief K B, Chen W, Shi Y, et al. The roadmap to 6G: AI empowered wireless networks. IEEE Commun Mag, 2019, 57: 84–90
262 Zhang N, Yang P, Ren J, et al. Synergy of big data and 5G wireless networks: opportunities, approaches, and challenges. IEEE Wireless Commun, 2018, 25: 12–18
263 Chen Z, Zhang Z, Yang Z. Big AI models for 6G wireless networks: opportunities, challenges, and research directions. IEEE Wireless Commun, 2024, 31: 164–172

264 Qian L, Zhu J, Zhang S. Survey of wireless big data. J Commun Inf Netw, 2017, 2: 1–18
265 Dai H N, Wong C W, Wang H, et al. Big data analytics for large scale wireless networks: challenges and opportunities. 2019. ArXiv:1909.08069
266 Chen T, Tang Q, Liu G. Efficient task scheduling and resource allocation for AI training services in native AI wireless networks. In: Proceedings of IEEE International Conference on Communications Workshops (ICC Workshops), Rome, 2023. 637–642
267 Chen Y, Li R, Zhao Z, et al. NetGPT: an AI-native network architecture for provisioning beyond personalized generative services. IEEE Netw, 2024, 38: 404–413
268 Challita U, Ryden H, Tullberg H. When machine learning meets wireless cellular networks: deployment, challenges, and applications. IEEE Commun Mag, 2020, 58: 12–18
269 Passalis N, Tefas A, Kanniainen J, et al. Deep adaptive input normalization for time series forecasting. IEEE Trans Neural Netw Learn Syst, 2020, 31: 3760–3765
270 Zhang H, Zhang Y F, Zhang Z, et al. LogoRA: local-global representation alignment for robust time series classification. IEEE Trans Knowl Data Eng, 2024, 36: 8718–8729
271 Xu C, Du X, Fan X, et al. FastVSDF: an efficient spatiotemporal data fusion method for seamless data cube. IEEE Trans Geosci Remote Sens, 2024, 62: 1–22
272 Fang C, Hu Z, Meng X, et al. DRL-driven joint task offloading and resource allocation for energy-efficient content delivery in cloud-edge cooperation networks. IEEE Trans Veh Technol, 2023, 72: 16195–16207
273 Qiao L, Mashhadi M B, Gao Z, et al. Latency-aware generative semantic communications with pre-trained diffusion models. IEEE Wireless Commun Lett, 2024, 13: 2652–2656
274 Shin M, Ma J, Mishra A, et al. Wireless network security and interworking. Proc IEEE, 2006, 94: 455–466
275 IMT-2030 (6G) Promotion Group. 6G AlaaS requirement research (white paper). 2023. https://www.imt2030.org.cn/html//default/en/Publications/Report/list-6.html?index=2
276 6GANA TG1. 6G AlaaS requirements whitepaper (in Chinese). 2023. https://www.6g-ana.com/upload/file/20231114/638355564809569589842750 6.pdf
277 Partarakis N, Zabulis X. A review of immersive technologies, knowledge representation, and AI for human-centered digital experiences. Electronics, 2024, 13: 269
278 Guo J, Chen H, Song B, et al. Distributed task-oriented communication networks with multimodal semantic relay and edge intelligence. IEEE Commun Mag, 2024, 62: 82–89
279 Benotsmane R, Kovács G, Dudás L. Economic, social impacts and operation of smart factories in Industry 4.0 focusing on simulation and artificial intelligence of collaborating robots. Soc Sci, 2019, 8: 143
280 Cui Q, Zhao X, Ni W, et al. Multi-agent deep reinforcement learning-based interdependent computing for mobile edge computing-assisted robot teams. IEEE Trans Veh Technol, 2023, 72: 6599–6610
281 Santos M A G, Munoz R, Olivares R, et al. Online heart monitoring systems on the internet of health things environments: a survey, a reference model and an outlook. Inf Fusion, 2020, 53: 222–239
282 Zhang S, Chen J, Lyu F, et al. Vehicular communication networks in the automated driving era. IEEE Commun Mag, 2018, 56: 26–32
283 Gao B, Liu J, Zou H, et al. Vehicle-road-cloud collaborative perception framework and key technologies: a review. IEEE Trans Intell Transp Syst, 2024, 25: 19295–19318
284 Yang Y, Wu J, Chen T, et al. Task-oriented 6G native-AI network architecture. IEEE Netw, 2024, 38: 219–227
285 Shen J, Wu B, Xiang W, et al. Novel bandwidth-aware network coding for fast cloud-of-clouds disaster backup. IEEE Trans Netw Serv Manage, 2025. doi: 10.1109/TNSM.2024.3524787
286 3rd Generation Partership Project (3GPP). Study on artificial intelligence (AI)/machine learning (ML) for NR air interface. TR 38.843. https://www.3gpp.org/ftp/Specs/archive/38_series/38.843
287 3rd Generation Partership Project (3GPP). Study on artificial intelligence (Al)/machine learning (ML) for mobility in NR. TR 38.744. https://www.3gpp.org/ftp/Specs/archive/38_series/38.744
288 3rd Generation Partership Project (3GPP). Study on further enhancement for data collection. TR 37.817. https://www.3gpp.org/ftp/Specs/archive/37_series/37.817
289 3rd Generation Partership Project (3GPP). Study on enhancements for artificial intelligence (AI)/machine learning (ML) for NG-RAN. TR 38.743. https://www.3gpp.org/ftp/Specs/archive/38_series/38.743
290 3rd Generation Partership Project (3GPP). 3GPP AI/ML consistency alignment. TR 22.850. https://www.3gpp.org/ftp/Specs/archive/22_series/22.850
291 5G-MoNArch. 5G mobile network architecture for diverse services, use cases, and applications in 5G and beyond. 2019. https://5g-monarch.eu/
292 6G Alliance of Network AI (6GANA). Whitepaper on distributed learning of 6G. 2023. https://www.6g-ana.com/upload/file/20240129/63842121412556678787856 30.pdf
293 Zhang R, Xiong K, Du H, et al. Generative AI-enabled vehicular networks: fundamentals, framework, and case study. IEEE Netw, 2024, 38: 259–267
294 Chen L, Qi J, Su X, et al. REMR: a reliability evaluation method for dynamic edge computing network under time constraint. IEEE Int Things J, 2023, 10: 4281–4291
295 Kourouklidis P, Kolovos D, Noppen J, et al. A model-driven engineering approach for monitoring machine learning models. In: Proceedings of ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C), Fukuoka, 2021. 160–164
296 Yao L, Ge Z. Industrial big data modeling and monitoring framework for plant-wide processes. IEEE Trans Ind Inf, 2021, 17: 6399–6408
297 Nan G, Li Z, Zhai J, et al. Physical-layer adversarial robustness for deep learning-based semantic communications. IEEE J Sel Areas Commun, 2023, 41: 2592–2608
298 Cai X, Shi K, Sun Y, et al. Stability analysis of networked control systems under DoS attacks and security controller design with mini-batch machine learning supervision. IEEE Trans Inform Forensic Secur, 2024, 19: 3857–3865
299 Xu P, Wang K, Hassan M M, et al. Adversarial robustness in graph-based neural architecture search for edge AI transportation systems. IEEE Trans Intell Transp Syst, 2023, 24: 8465–8474
300 Pu C, Wang K, Xia Y. Robustness of link prediction under network attacks. IEEE Trans Circ Syst II, 2020, 67: 1472–1476
301 Dong C, Xiong X X, Xue Q L, et al. A survey on the network models applied in the industrial network optimization. Sci China Inf Sci, 2024, 67: 121301
302 Yang R, Zhang Z, Zhang X, et al. Meta-learning for beam prediction in a dual-band communication system. IEEE Trans Commun, 2023, 71: 145–157
303 Xu Z, Wang S, Zhang Y J A. SAMBA: scenario-adaptive meta-learning for mmWave beam alignment. In: Proceedings of IEEE Globecom Workshops (GC Wkshps), Kuala Lumpur, 2023. 1–6
304 Kim H, Choi J, Love D J. Massive MIMO channel prediction via meta-learning and deep denoising: is a small dataset enough? IEEE Trans Wireless Commun, 2023, 22: 9278–9290
305 Wu B, Zou S, Liwang M, et al. Explainable application intent for zero-touch networking: an incorporation of hypergraph

and transformer. IEEE Trans Commun, 2025. doi: 10.1109/TCOMM.2025.3529260
306 Zou S, Liwang M, Wu B, et al. Intent-oriented network slicing with hypergraphs. IEEE Netw, 2024. doi: 10.1109/MNET.2024.3454117
307 Nichol A, Achiam J, Schulman J. On first-order meta-learning algorithms. 2018. ArXiv:1803.02999
308 Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res, 2008, 9: 2579–2605
309 Wang R, Yang L, Tang T, et al. Robust federated learning for heterogeneous clients and unreliable communications. IEEE Trans Wireless Commun, 2024, 23: 13440–13455
310 Zheng P, Zhu Y, Hu Y, et al. Federated learning in heterogeneous networks with unreliable communication. IEEE Trans Wireless Commun, 2024, 23: 3823–3838
311 Pang Y, Zhang H, Deng J D, et al. Collaborative learning with heterogeneous local models: a rule-based knowledge fusion approach. IEEE Trans Knowl Data Eng, 2024, 36: 5768–5783
312 Liu X, Wang G, Liu Z, et al. Hierarchical reinforcement learning integrating with human knowledge for practical robot skill learning in complex multi-stage manipulation. IEEE Trans Automat Sci Eng, 2024, 21: 3852–3862
313 Yan Y, Tong X, Wang S. Clustered federated learning in heterogeneous environment. IEEE Trans Neural Netw Learn Syst, 2024, 35: 12796–12809
314 Wang Y, Wu Y, Chen X, et al. Incentive-aware decentralized data collaboration. In: Proceedings of the ACM on Management of Data, New York, 2023. 1–27
315 Liu Y J, Feng G, Sun Y, et al. Resource consumption for supporting federated learning enabled network edge intelligence. In: Proceedings of IEEE International Conference on Communications Workshops (ICC Workshops), Seoul, 2022. 1–6
316 Zhang R, Pan C, Wang Y, et al. Federated deep reinforcement learning for multimedia task offloading and resource allocation in MEC networks. IEICE Trans Commun, 2024, E107-B: 446–457
317 Ji Z, Qin Z, Tao X. Meta federated reinforcement learning for distributed resource allocation. IEEE Trans Wireless Commun, 2024, 23: 7865–7876
318 Cherif N, Jaafar W, Yanikomeroglu H, et al. RL-based Cargo-UAV trajectory planning and cell association for minimum handoffs, disconnectivity, and energy consumption. IEEE Trans Veh Technol, 2024, 73: 7304–7309
319 Samsi S, Zhao D, McDonald J, et al. From words to watts: benchmarking the energy costs of large language model inference. In: Proceedings of IEEE High Performance Extreme Computing Conference (HPEC), Boston, 2023. 1–9
320 Zawish M, Dharejo F A, Khowaja S A, et al. AI and 6G into the metaverse: fundamentals, challenges and future research trends. IEEE Open J Commun Soc, 2024, 5: 730–778
321 Zawish M, Ashraf N, Ansari R I, et al. Energy-aware AI-driven framework for edge-computing-based IoT applications. IEEE Int Things J, 2023, 10: 5013–5023
322 Li X, Zhang H, Shen Y, et al. Intelligent traffic data transmission and sharing based on optimal gradient adaptive optimization algorithm. IEEE Trans Intell Transp Syst, 2023, 24: 13330–13340
323 Bian J, Arafat A A, Xiong H, et al. Machine learning in real-time Internet of Things (IoT) systems: a survey. IEEE Int Things J, 2022, 9: 8364–8386
324 Chu K F, Lam A Y S, Li V O K. Traffic signal control using end-to-end off-policy deep reinforcement learning. IEEE Trans Intell Transp Syst, 2022, 23: 7184–7195
325 Wozniak M, Silka J, Wieczorek M, et al. Recurrent neural network model for IoT and networking malware threat detection. IEEE Trans Ind Inf, 2021, 17: 5583–5594
326 Mao Y, You C, Zhang J, et al. A survey on mobile edge computing: the communication perspective. IEEE Commun Surv Tutorials, 2017, 19: 2322–2358
327 Xiao D, Chen S, Ni W, et al. A sub-action aided deep reinforcement learning framework for latency-sensitive network slicing. Comput Netws, 2022, 217: 109279
328 Wang Y, Sun T, Li S, et al. Adversarial attacks and defenses in machine learning-empowered communication systems and networks: a contemporary survey. IEEE Commun Surv Tut, 2023, 25: 2245–2298
329 Jing X, Yan Z, Pedrycz W. Security data collection and data analytics in the Internet: a survey. IEEE Commun Surv Tut, 2019, 21: 586–618
330 Li K, Zheng J, Ni W, et al. Biasing federated learning with a new adversarial graph attention network. IEEE Trans Mobile Comput, 2025, 24: 2407–2421
331 Raja A, Njilla L, Yuan J. Adversarial attacks and defenses toward AI-assisted UAV infrastructure inspection. IEEE Int Things J, 2022, 9: 23379–23389
332 Wei K, Li J, Ding M, et al. Federated learning with differential privacy: algorithms and performance analysis. IEEE Trans Inform Forensic Secur, 2020, 15: 3454–3469
333 You F, Yuan X, Ni W, et al. Privacy-preserving multi-agent deep reinforcement learning for effective resource auction in multi-access edge computing. IEEE Trans Cogn Commun Netw, 2024. doi: 10.1109/TCCN.2024.3499342
334 Wang F, Xie M, Tan Z, et al. Preserving differential privacy in deep learning based on feature relevance region segmentation. IEEE Trans Emerg Top Comput, 2024, 12: 307–315
335 Lagendijk R, Erkin Z, Barni M. Encrypted signal processing for privacy protection: conveying the utility of homomorphic encryption and multiparty computation. IEEE Signal Process Mag, 2013, 30: 82–105
336 Valina L, Teixeira B, Reis A, et al. Explainable artificial intelligence for deep synthetic data generation models. In: Proceedings of IEEE Conference on Artificial Intelligence (CAI), Singapore, 2024. 555–556
337 Liu J, Zhao Y. Role-oriented task allocation in human-machine collaboration system. In: Proceedings of the 4th International Conference on Information Systems and Computer Aided Education (ICISCAE), Dalian, 2021. 243–248
338 Alam S, Khan M F. Enhancing AI-human collaborative decision-making in Industry 4.0 management practices. IEEE Access, 2024, 12: 119433
339 Dubois C, Le Ny J. Adaptive task allocation in human-machine teams with trust and workload cognitive models. In: Proceedings of IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, 2020. 3241–3246
340 Zhang X, Ke Q, Zhao X. Travel demand forecasting: a fair AI approach. IEEE Trans Intell Transp Syst, 2024, 25: 14611–14627
341 Pasricha S. AI ethics in smart healthcare. IEEE Consumer Electron Mag, 2023, 12: 12–20
342 Chen X, Dai W, Ni W, et al. Augmented deep reinforcement learning for online energy minimization of wireless powered mobile edge computing. IEEE Trans Commun, 2023, 71: 2698–2710
343 Sutton G J, Zeng J, Liu R P, et al. Enabling technologies for ultra-reliable and low latency communications: from PHY and MAC layer perspectives. IEEE Commun Surv Tut, 2019, 21: 2488–2524
344 Wang Y, Sun T, Yuan X, et al. Minimizing adversarial training samples for robust image classifiers: analysis and adversarial example generator design. IEEE Trans Inform Forensic Secur, 2024, 19: 9613–9628
345 Huang H, Duan L, Li C, et al. A secure and lightweight aggregation method for blockchain-based distributed federated learning. In: Proceedings of IEEE International Conference on Web Services (ICWS), 2024. 447–456
346 Wu Q, Zhang R. Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming. IEEE Trans Wireless Commun, 2019, 18: 5394–5409
347 Zhang N, Zhang J F, Xing C W, et al. Intelligent secure near-field communication. Sci China Inf Sci, 2024, 67: 199302

348 Zhu G X, Lyu Z H, Jiao X, et al. Pushing AI to wireless network edge: an overview on integrated sensing, communication, and computation towards 6G. Sci China Inf Sci, 2023, 66: 130301
349 Kurunathan H, Huang H, Li K, et al. Machine learning-aided operations and communications of unmanned aerial vehicles: a contemporary survey. IEEE Commun Surv Tut, 2024, 26: 496–533
350 DeepSeek-AI, Liu A, Feng B, et al. Deepseek-V3 technical report. 2024. ArXiv:2412.19437
351 Al-Hraishawi H, Alsenwi M, Ur Rehman J, et al. Digital twin for enhanced resource allocation in 6G non-terrestrial networks. IEEE Commun Mag, 2025, 63: 47–53
352 Abbas K, Nauman A, Bilal M, et al. AI-driven data analytics and intent-based networking for orchestration and control of B5G consumer electronics services. IEEE Trans Consumer Electron, 2024, 70: 2155–2169
353 Duan J, Yu S, Tan H L, et al. A survey of embodied AI: from simulators to research tasks. IEEE Trans Emerg Top Comput Intell, 2022, 6: 230–244
354 Wang Y, Ni W, Yi W, et al. Federated contrastive learning for personalized semantic communication. IEEE Commun Lett, 2024, 28: 1875–1879

Appendix A

Table A1 List of abbreviations.

Abbreviation	Definition	Abbreviation	Definition
3GPP	3rd generation partnership project	LLM	Large language model
4G	Fourth-generation	LoS	Line-of-sight
5G	Fifth-generation	LAM	Large-scale AI model
6G	Sixth-generation	ML	Machine learning
AI	Artificial intelligence	MIMO	Multiple-input multiple-output
AI4NET	AI for network	mMTC	Massive machine type communications
AIaaS	AI as a service	MEC	Mobile edge computing
AR	Augmented reality	MARL	Multi-agent reinforcement learning
BS	Base station	NLP	Natural language processing
CSI	Channel state information	NLoS	Non-line-of-sight
CNN	Convolutional neural network	NWDAF	Network data analytics function
CN	Core network	NTN	Non-terrestrial network
CT	Core network and terminal	NR	New radio
CP	Cyclic prefix	NET4AI	Network for AI
CPU	Central processing units	O&M	Operation and maintenance
DL	Deep learning	OFDM	Orthogonal frequency division multiplexing
DRL	Deep reinforcement learning	O-RAN	Open radio access networks
DT	Digital twins	PFL	Personalized federation learning
DTC	Digital twins channel	QoS	Quality of service
DQN	Deep Q-network	QoAIS	Quality of AI service
DDPG	Deep deterministic policy gradient	RL	Reinforcement learning
DNN	Deep neural network	RAN	Radio access network
eMBB	Enhanced mobile broadband	REK	Radio environment knowledge
ETSI	European telecommunication standardization institute	RNN	Recurrent neural network
ENI	Experiential networked intelligence	SON	Self-organizing networks
FL	Federated learning	SGCS	Square of the generalized cosine similarity
FedAvg	Federated averaging	SINR	Signal-to-interference-plus-noise ratio
gNB	Generation node B	SNR	Signal-to-noise ratio
GPU	Graphics processing units	SL	Split learning
HARQ	Hybrid automatic repeat request	SMPC	Secure multi-party computation
IP	Internet protocol	SI	Semantic information
ITU	International telecommunication union	SA	System aspects
ICT	Information and communication technology	TSG	Technical specification groups
INC	In-network computing	uRLLC	Ultra-reliable low latency communications
IMT	International mobile communications	UPF	User plane functions
IRS	Intelligent reflecting surface	UAV	Unmanned aerial vehicle
ISAC	Integrated sensing and communication	VR	Virtual reality
IoT	Internet of Things	WG	Working groups
KPI	Key performance indicator	XR	Extensive reality
LEO	Low earth orbit		