Enabling Mobile AI Agent in 6G Era: Architecture and Key Technologies

Ziqi Chen, Qi Sun, Nan Li, Xiang Li, Yang Wang, and Chih-Lin I

Abstract

With the advent of mobile networks, we are witnessing an unprecedented shift in the landscape of mobile network services, evolving from traditional voice calls to advanced artificial intelligence (AI) services. This paper delves into the intricacies of this evolution, particularly emphasizing the deep integration of AI agents into 6G networks. Despite recent researches in using large language model (LLM) and AI agent for network automation, the fundamental mobile AI agent use cases, their network requirements, potential network architecture and enabling technologies for supporting the pervasive AI agents in 6G era are largely unexplored. In this article, we present an in-depth analysis of typical mobile AI agent use cases in 6G, consisting of AI agent-based 6G network automation, handheld personalized agents, connected robotics and autonomous systems, and wearable AI agent. Then, we elucidate a novel system architecture that supports identified use cases. The article also addresses core aspects of enabling technologies, including 6G agent and application agent collaboration, efficient model and memory management, coordinated agent-to-agent communication and support of multi-modal data transmission. A proof of concept prototype is also presented to demonstrate 6G agent and application AI agent collaboration. Finally, three challenges and research directions: energy saving, security protection and AI agent tailored communication are discussed. This article lays a foundation for understanding the role of 6G in realizing the full potential of AI agents in various applications.

Introduction

Recent advances in LLM, such as Generative Pretrained Transformer (GPT) has already significantly changed people's everyday life, and totally altered the way people perceive AI's role in the future. Following such development momentum, we are about to see the emergence of Artificial General Intelligence (AGI) in near future, penetrating in everyone's daily life, and creating an AI-driven society with AI-empowered applications from personal assistant to autonomous vehicles, from entertainment to manufacturing. To realize AGI, LLM alone is not enough due to its limited private or personalised knowledge, limited memory, logical reasoning, evaluation, and refinement abilities. Therefore, scientists have proposed AI agents as key enabler of AGI, which utilize LLM as central controller, and equip it with profiling module, memory module, planning module and action module [1].

The telecommunication industry is currently researching on leveraging LLM and AI agents to achieve higher level of network automation, such as automatic orchestration and maintenance (OAM) and intent-based networking (IBN), where network OAM personnel simply issues network configuration intents and let AI agent translate them into network tasks and run execution-feedback-adjust-monitoring process automatically [2]. Work has also emerged to discuss potential of collaborative cloud-edge methodology for personalized generative services [3], Cloud-Edge-Mobile device collaborative full life-cycle management of LLMs [4], and constructing domain-adapted LLMs for network planning tasks [5].

At the same time, we have observed that mobile network itself is extending its capabilities beyond the communication services. Transformed by data, operation, information, and communication technologies (DOICT) convergence, next generation mobile networks are expected to provide integrated AI and communication, and integrated sensing and communication services [6]. As a result, networks are transforming into a communication and computing integrated platform [7], [8], offering opportunities for the mobile network to provide edge AI agents with multifaceted services beyond simply data transmission.

However, in future AI-driven society, mobile network will face the emergence of large number of AI agents of quite various form factors, such as handheld AI agents, wearable AI agents, robotic AI agents, vehicle AI agents. Thus, a critical, yet unanswered question persists: what role will next-generation mobile networks play in an AI-driven society? As far as we know, efforts on understanding future AI agent use cases, its functional requirements to 6G network, its enabling technologies and open issues remain in a nascent stage. In this article, we first analyse four typical AI agent use cases closely related to 6G, namely AI agent based 6G network automation, handheld personalized agents, connected robotics and autonomous systems, and wearable AI agents. Then, a network architecture to fulfil AI agent service requirements for 6G is presented, consisting of four layers: service exposure layer, AI

Digital Object Identifier: 10.1109/MNET.2024.3422309
Date of Current Version: 16 September 2024
Date of Publication: 15 July 2024

Ziqi Chen, Qi Sun, Xiang Li, Yang Wang, and Chih-Lin I are with the China Mobile Research Institute, Beijing 100053, China; Nan Li (corresponding author) is with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China, and also with the China Mobile Research Institute, Beijing 100053, China.

0890-8044/24©2024IEEE

agent layer, function layer, and infrastructure layer. Furthermore, four enabling technologies, including 6G agent and application agent collaboration, efficient model and memory management, coordinated agent-to-agent communication and multi-modal data transmission are discussed. Finally, challenges and research directions like energy saving, security protection and AI agent tailored communication were discussed followed by overall conclusion.

AI Agents in 6G

Over recent decades, mobile communication technologies have undergone transformative changes. With the rapid development of the AI technology, AI agents is envisioned to be pervasive in our daily life. In Fig. 1, we listed some typical example of AI agent entities depending on computing capability and interaction level with human and environment. These AI agent entities will present in various locations, relying on 6G network on large number of use cases, which are presented in Fig. 2, ranging from house/office to open public environments to vertical industries. In this context, the 6G network is actively embracing the new technology trends on AI agents, and 6G also emerges as an essential force in the AI agent's era. In this section, we will first briefly discuss how AI agent could enhance network automation. Then, we illustrate the critical roles 6G networks play in enhancing AI agent capabilities by investigating three use cases, each representing one typical categories of the AI agents, from augmenting hand-held AI agents, to assisting connected robotics and autonomous systems, until offloading whole AI agent computing for wearable native AI devices.

6G AI Agent for 6G Network Automation

To achieve the autonomous mobile network vision, the 6G network could employ an AI agent as its own controller. To be more specific, The AI agent in 6G network has following roles:

Intent-Based Network Management: The 6G AI agent enhances network management by translating high-level intents from operators into specific configurations and actions, deploys them within the network infrastructure, enabling dynamic and responsive management without detailed manual intervention. It uses various network tools to predict and prevent equipment failures by scheduling maintenance or self-checks during low-traffic periods, and automatically lookup device log and generate recovery methods for minimal downtime. Additionally, the AI agent can even automatically generate function-oriented code or programs to deal with network problem, allowing for more prompt and versatile solutions to various network circumstances.

Automated Network Optimization: Utilizing its network tools within 6G network, AI agent can continuously monitor and analyze network performance in real-time, identifying potential issues and optimizing configurations proactively. This includes load balancing, SLA assurance, and interference management. Recent researches on time-serious prediction with LLM also gives AI agent native capabilities to predict network KPIs, like congestion or service degradation, even without using traditional machine learning algorithm. This give AI agent more responsive abilities to take preemptive actions to maintain optimal network performance and quality of service (QoS).

Handheld Personalized Agents

Recent advancements in mobile System on Chip (SoC) technology have enabled vendors to develop chips capable of running LLMs locally on handheld devices. This development allows users to access LLM functionalities offline, enhancing data privacy, albeit with some compromises in LLM capabilities and device battery life. In the 6G

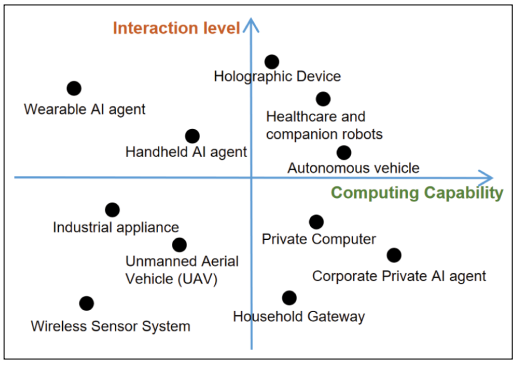

FIGURE 1. AI agent examples, positioned by their computing capability and interaction level.

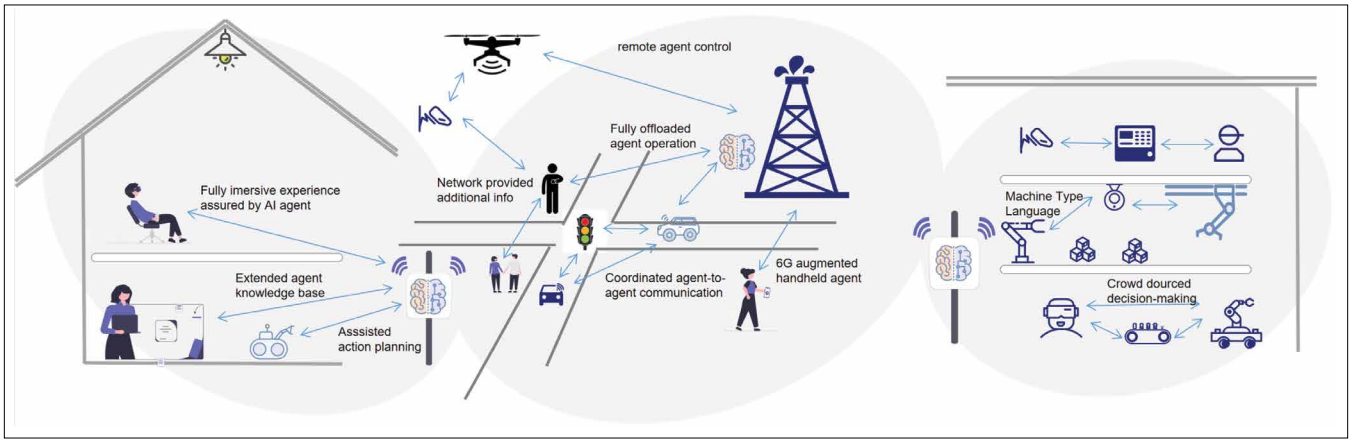

FIGURE 2. AI agents in 6G typical use cases.

era, personalized AI agents on handheld devices will become widespread, with LLMs running on the device SoC as agent core, further emphasizing the critical role of 6G networks in supporting personalized AI agents on handheld devices in the following aspects.

Enrichment Information provision: Handheld AI agents could benefits from 6G network in terms of dedicated enrichment information which helps enhancing AI agent knowledge about and around the user, providing the phone's AI agent with insights that are critical for complex user queries.

Utilizing 6G's unique user information that are not available on handheld devices, this enrichment information can include user tailored information like real-time traffic in proximity, local weather updates, news that concerns the user or emergency alarms. 6G network could also proactively provide handheld AI agents with important information about the user, like predicting when the user will arrive a home, or when the user will experience lowered internet connection speed. Furthermore, Handheld AI agent are offered with additional functions that 6G network could help provide, like configuring unique sensing objects, controlling remote IoT devices or coordinating other AI agents to jointly complete a task.

Computing resource provision: Limited by handheld device's computing and storage capability, as well as battery capacity, many of AI agent operation needs aiding of additional computing resources. 6G network, which aims to provide AI and communication integrated services, will help handheld AI agents in memory storage, model fine-tuning, model updates and more.

Longer the user uses, the larger AI agent's multi-modal memory size will be. Fortunately, many memories are location, time or context correlated, which means 6G networks, while providing basic communication services, could efficiently cache AI agent's memory, and only download it to handheld devices when certain events are triggered, like the user entered an specific area or asked about it. Fine-tuning is a technique that uses new data to partially train a LLM without largely modifying the model weights. This is needed as AI agent evolve itself to better understand the user. With much larger computing resources, AI agents could request 6G network to fine-tune its model, and 6G network should coordinate the model and data uplink and downlink transmission as well as model training jointly considering communication and computing resources. lastly, AI agent models are expected to get version updates just like software, then distributed to handheld devices by agent vendors from cloud. As new versions may lack user personalization, before deploying, 6G network will support distributed and collaborative model training between cloud, edge and UEs.

Connected Robotics and Autonomous Systems

Autonomous robots, integral to modern technological advancements, is witnessing a boom in its fast-growing capability range. As a consequence, more and more autonomous robots are expected to be integrated into the everyday life to provide various services [9]. These robots encompass diverse types such as autonomous vehicles, unmanned aerial vehicles (UAVs), service and industrial robots, healthcare and companion robots, etc.

Some robots are assigned with specific job, while others are designed to serve general capability. The complexity of robot tasks necessitates sophisticated decision-making, positioning AI agents as central components in many robotic systems.

Efficient communication is vital for these robots. Direct robot-to-robot communication ensures smooth collaboration. For example, in a factory, multiple robots might work together on an assembly line and communicate to coordinate tasks like welding, painting, and assembly, ensuring a smooth workflow; self-driving cars communicate with each other to share traffic information, road conditions, and to coordinate movements. In disaster recovery missions, different robots (like drones and ground robots) share data about the environment, victim locations, and obstacles, aiding in effective rescue operations. Despite advancements, current communication systems often fall short in supporting the intricate requirements of such sophisticated robotic networks.

Common context building: Context information represents a type of information that possess the necessary knowledge that an AI agent needs to know to successfully complete the task. It is extracted from raw data, abstracted to a level that helps AI agents to understand and adapt to their dynamic environments while moving.

Context information is often position based, such as local environment information, nearby other agents and their status, etc. For example, cooperative perception techniques, like using SLAM to jointly build high-resolution 3D map relies on fast fusion of high definition local and remote maps collected by participating AI agents. 3D map as context information is particularly crucial for agents involved in navigation, like autonomous vehicles or drones, where real-time environmental understanding is key for safety and efficiency. As surrounding context information keep updating frequently, 6G, connecting those AI agents, are expected to help generate, store, load, update and distribute context information to assisted AI agents.

Collaborative communication assurance: It is expected that future industry robots will embed a large variety of sensors that generate huge amount of data to process. Each type of sensing data requires a different AI/ML model. For resource and energy efficiency, these processing tasks could be offloaded to other AI agents, or 6G network infrastructures in proximity, as long as QoS is satisfied. Also, collaborative inference among AI agents needs transmitting intermediate data and model layers between nearby AI agents. The AI agents will employ a variety of communication schemes to satisfy complex task automation. For example, AI agent may simply issues its data type, QoS, and destination to 6G system, and 6G system should find best ways to send them to target AI agent whether using direct or indirect communications, or employing both. Therefore, the selection and reselection of relaying AI agents, dedicated communication resources, and data QoS adaptive modification

capabilities are expected, requiring higher level of information and capability exposure between 6G network and AI agents.

Wearable AI Agent

With the rapid advances of LLM and other generative AI applications, we have witnessed a new type of devices being introduced into the market: wearable native AI device. This kind of device varies in their form factor. There is a big LLM running behind each wearable device, allowing them to function as personal assistant, interacting with user in more natural ways than handheld devices. Normally, these devices can see, hear external environments, and talk to user or display images in front of users. For portability, they normally rely on consistent internet access and cloud computing to run the LLM. To further enhance their capability, AI agents will gradually replace LLMs and become the operation entity behind wearable AI devices, utilizing LLM and other tools to fulfil the tasks that LLMs are weak on. In 6G, the network is expected support wearable devices' consistent and stable internet connection, offloading the agent operation, and allows for responsive and high-speed multi-modal data transmission.

Fully offloaded agent operation: The 6G network can effectively offload the operations of wearable AI agent devices by providing a robust infrastructure for AI agents and LLMs to run within the network, reducing the processing load on the wearable devices. Furthermore, unlike handheld devices, wearable devices do not have any application on device. So, to support broad AI agent capabilities just like handheld AI agents, 6G should provide diverse commonly used application, potentially through API tools. The tools could include search engines, calculator, calendar queries, smart home control, schedule management, health data management, account authentication workflow and more.

Responsive multi-modal interaction: Wearable AI agents represent a type a device that provides higher degree of interaction level with human. They can receive information through filming the environment, recording the sounds, or even detect human thinking through Brain-Computer interfaces [10]. They can also generate multi-modal outputs, like fully immersive XR experience as well as real-time tactile and haptic information. Therefore, wearable communication needs multi-modal data transmission of higher speed, lower delay in dynamic environments (e.g., higher user movement speed). The synchronization of associated data flows is also essential to improve user's sense of presence and realism. To further complicate the scenario, wearable may have limited antenna number, lower transmitting power and reduced networking capability than handheld devices, posing even more challenging network requirements.

Referencing 3GPP SA technical reports in R18 and R19 [11], [12], [13], we have conjured a requirement table of above discussed use cases, as shown in Fig. 3.

System Architecture

To fully support above identified AI agent use cases, we present an 6G system architecture that is centred by AI agents. In the architecture, 6G network employs the 6G AI agent as its own 6G network intelligence engine for enabling autonomous network. The 6G network system architecture depicted encompasses 4 layers: service exposure layer, agent layer, function layer and infrastructure layer, as shown in Fig. 4.

The Service Exposure Layer in the 6G network architecture functions as the interface between the

Use Case	Functional Requirement	Network KPI Requirements				
		operation	max allowed end-to-end latency	message size	data rate	communication service availability
Handheld Personalized Agents	• Collect and tailor user-specific enrichment information • Alleviate Handheld AI computational burdens • Dynamically configure comm. and comp. resource for AI agent updates	transmitting 3D object recognition model to UE (AlexNet)	1s	240 MByte	1.92 Gbit/s	99.9 %
		transmitting LLM base model to UE (LLAMA-3-8B)	5s	16 GByte	25.6 Gbits/s	99.9 %
Connected Robotics and Autonomous Systems	• Building and distributing context information • Assure efficient and reliable direct device connection	Haptic information (position, velocity) sharing-low dynamic robotics	5-10ms	n DoFs: (2n)-(8n) (n=1, 3, 6) (Byte)	0.8 - 200 kbit/s (with compression)	99.999%
Wearable AI Agents	• Support real-time multi-modal communication for interactive feedback • Ensure synchronous arrival of associated data transmission for associated data flows • Provide running platform for AI agent and needed tools/models	immersive multi-modal VR (video DL)	10 m	1500 Byte	1-100 Mbit/s	99.9%
		immersive multi-modal VR (audio DL)	10ms	50 Byte	5-512 kbit/s	99.9%
		immersive multi-modal VR (Haptic feedback)	5ms	1 DoF: 2-8 3 DoFs: 6-24 6 DoFs: 12-48 (Byte)	16 kbit/s -2 Mbit/s (without haptic compression encoding);	99.9%
		audio-visual-tactile data flow synchronozation	audio-tactile		audio delay: 50 ms tactile delay: 25 ms	
			visual-tactile		visual delay:15 ms tactile delay: 50 ms	

FIGURE 3. Requirements of typical use cases.

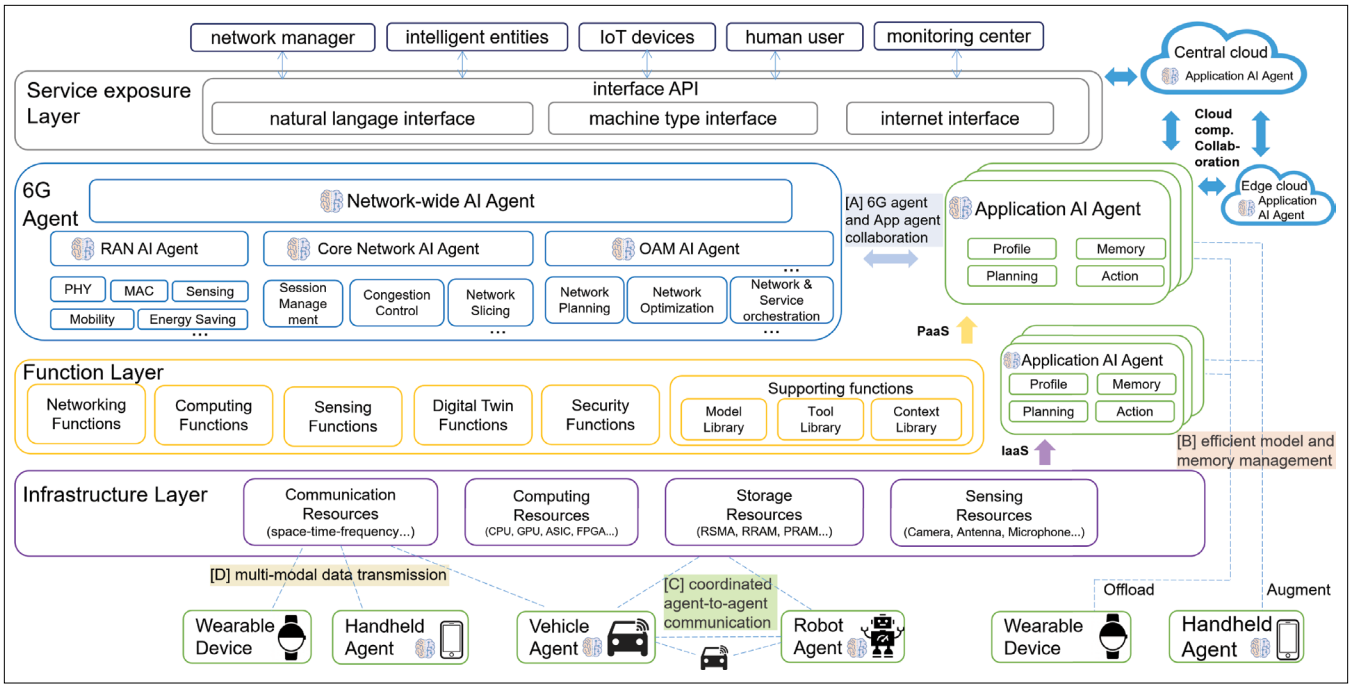

FIGURE 4. Proposed architecture of AI agent-driven 6G network.

network and its users, whether they are machines or humans. It's here that the network's services are made accessible to various entities through a set of defined APIs. These APIs are tailored for different types of interactions: natural language interfaces for human interaction, machine-type interfaces for device communication, and internet interfaces for broader connectivity.

The Agent Layer in the 6G network system architecture houses various specialized agents:

- 6G AI agent: Acts as the powerful intelligent and automation entity for 6G network, managing communications and network-specific tasks. 6G agent has multifaceted components. Inspired by [14], the 6G agent is orchestrated by 3 layers of AI agent. Positioned at the top is network-wide model agent. Positioned at the middle are AI agents for different technical domains, such as Radio Access Network (RAN), Core Network (CN), or OAM. Positioned at the bottom are AI agents focusing on more specialized scenarios, for example, agents for physical layer of RAN and agents for network optimization of OAM.
- Application AI agent: Can also exist in this layer, leveraging Infrastructure as a Service (IaaS) or Platform as a Service (PaaS) provided by infrastructure layer or function layer of the 6G network. They may also be deployed in central cloud, edge cloud or mobile devices. Application AI agent can be categorized into two types: user application agent and functional application agent. The user application agents are offloaded AI agents to 6G by user, offering minimum viable capabilities, for example the AI agents running behind wearable devices. Functional application agents are targeted at solving a particular kind of task to other user application agents, offering extendable capabilities on top of user application agents, for example the AI agents that generate 3D-map and coordinate routes for vehicles using aggregated sensor data and vehicle status (e.g. speed, location, intention). This kind of decoupled design of basic user application agent and extensible functional application agent allows for more flexible AI agent capability evolution.

6G AI agent manages the network operation, and also interact with application agents to provide enrichment information, or receive network control intents for assuring better user experience.

The Function Layer in the 6G network architecture is where core operational capabilities reside. It is composed of 6 essential functional modules:

- Networking Functions: These are responsible for managing data transmission and network connectivity.
- Computing Functions: This involves processing and computational functions necessary for network operations and service delivery.
- Sensing Functions: These are used for collecting environmental data, which is crucial for providing sensing and communication integrated services.
- Digital Twin Functions: These are used for simulating network configurations, predicting network performances and providing virtual spaces for training new algorithms.
- Security Functions: These are responsible for managing multi-level security issues, such as user authentication, data encryption, key management and fail-safe protocol.
- Supporting Functions: These are responsible for providing a running environment for AI agents, including necessary models, tools and contexts libraries.

This layer is foundational to the network's operation, and also providing the necessary network AI service and capabilities for the application

agents to execute their tasks effectively by offering PaaS.

The Infrastructure Layer in the 6G network architecture is the foundational level that allocates and manages various infrastructure resources critical to network functionality:

- Communication Resources: Pertains to the space-time-frequency resources essential for wireless communication.
- Computing Resources: Includes the processing power required for network operations, services, and AI agent functionalities.
- Storage Resources: Involve various advanced storage technology that underpin the network's infrastructure, potentially incorporating computing in memory technology combining the two kind of resources together.
- Sensing Resources: Encompasses the sensors and data acquisition systems necessary for gathering information from the environment.

This layer enables IaaS to application agents, providing a versatile environment for AI agents who could access 6G network computing and communication resources.

6G network could also leverage central cloud and edge cloud computing resources to allow for more versatile AI agent deployment, or further enhance application agent capability.

Below 6G network architecture, various forms of AI agent entities are inter-connected through direct or indirect connections provided by 6G. The architecture view shows important enabling technologies to realize above mentioned use cases, respectively [A]: 6G agent and application agent collaboration, [B]: efficient model and memory management, [C]: coordinated agent-to-agent communication, and [D]: multi-modal data transmission.

In the proposed architecture, offloading some application AI agent's computing inside 6G network shows benefits to reaching carbon footprint goal as a whole. The pooling gain of running unified foundation model in 6G network for many mobile users means an increase of computing hardware and energy utilization. Additionally, 6G network infrastructures, potentially consisted of cloud-based heterogeneous hardware, provides low energy cost LLM inferences utilizing more advanced computing architecture, more sustainable energy source, avoiding running energy consuming LLMs on User Equipment distributively for unnecessarily long time.

Enabling Technologies

6G Agent and Application Agent Collaboration

In the proposed 6G network architecture, the 6G agent and application agent work in concert to enable more personalized functionalities and services. For example, user location information provided by 6G agent could help user application agent understand current environment, user habit information could help user application agent understand the user more deeply, sensed environment data provided to user application agent becomes its eye and ear, comprehending the surrounding. The clock synchronization accuracy related information of multiple UEs (e.g., autonomous mobile robots) provided by 6G agent are useful for determining the suitable collaborative robots and their corresponding control action.

6G agent could also help to customize the network capability, and provide on-demand deterministic connection and computing services to the user application AI agents. It is more desired in vertical industry use cases, as the communication links in it is quiet heterogeneous, and each may have quiet different QoS in dynamic environment. Take the smart factory scenario as an example, diversified services such as camera surveillance, automated visual inspection in smart production line, and collaborative robots (cobots) for automated product delivery, may coexists with differentiated traffic patterns, connection latency, reliability requirements and service priorities. Enrichment information of the services such as the traffic packet periodicity, the priority of the service data flow, the synchronization level of multiple cobots, as well as the UE level or finer granularity of the data flow or even IP packet level communication requirements can be provided by the user application AI agent to the 6G agent. So that 6G agent can autonomously optimize the network behavior to better suit the user application AI agents' need and potentially improve the resource and energy efficiency.

Additional functional application agents can be flexibly integrated into 6G system, leveraging 6G edge computing infrastructure to provide task-specific services to other AI agents, either deployed within 6G system or running on wireless devices. Furthermore, Through collaborating with 6G agent, functional application agents could get helpful information about the network as well as ask 6G agent to execute network related tasks to facilitate the completion of functional tasks. For example, in road intersections a traffic scheduling functional application agent is deployed in a 6G node, serving other vehicles AI agents passing by, helping constructing area context information, assuring prompt message sending and facilitating auto-driving through crowd-sourced decision making.

Efficient Model and Memory Management

Traditionally, edge caching is the way to deliver low-latency content. However, in 6G era where AI agent services is ubiquitously needed, caching does not only mean storing popular contents at edges. Rather, it is the agent model and its personalized memory that needs delicate treating in 6G networks.

As agent models are often LLMs, the model sizes are very large, containing billions to trillions of parameters. Frequent transmission of such large files poses stringent requirements on 6G network. To tackle this problem, various techniques can be employed to reduce the model size needed to be transmitted. Quantization is the technique to transform model weights from higher-precision numerical representations into lower-precision formats, e.g., 8-bit or 4-bit. Sparsification is the technique to reduce the number of active parameters by making many of its weights zero, or pruning certain neurons or layer. These two techniques, among other size shrinking methods, can be applied in the network to make LLMs not only more efficient to transmit, but also easier to run in mobile devices, but at the cost of model accuracy. Instead of transmitting the whole model, transmitting "add-on" models is more efficient. Modern fine-tunning methods, such as Low Rank

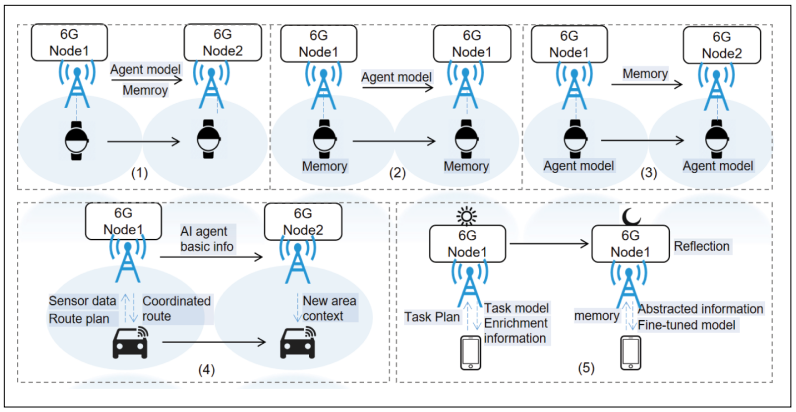

FIGURE 5. Examples of model and memory management scenarios.

Adaptation, trains a separate model of the same structure to the original model, but with much smaller parameter size (low-rank weight matrix). The trained model can be merged into the original model easily to produce a fine-tuned model dedicated specific dataset. Therefore, by only transmitting those trained "add-on" model, and performing the merging process at the receiver, a large proportion of bandwidth can be saved.

Memory is the major differentiator that differs AI agent from normal LLM. Currently there are 3 ways for LLM-based AI agent to store its memory. User prompts and agent responses, like chatting history, could be put in the context of LLM, which allows for LLM in-context learning, equivalent to letting LLM memorizing temporary knowledge. However, this kind of memory storage is poised to be forgotten when chatting history exceeds context window. An embedding vector database is a specialized database designed to store and manage high-dimensional vector representations (embedding) of data. Embedding refers to the process of transforming data into a fixed-size vector of continuous numbers, capturing the semantic meaning of the original data. With Retrieval Augmented Generation (RAG), LLM is able to rapidly query and retrieve memory from external storage (e.g. embedding vector database) using similarity search tools like Facebook AI Similarity Search (FAISS). This way of memorizing is long-term, and easy to manage and manipulate, e.g. adding, replacing and deleting certain memory is easy. However, as the memory is still relying on external database, its accessing speed is limited, and hard for agent to acquire new internal abilities through external memory. The last method, fine-tuning, allows using memory to re-train the LLM parameters in small scales, generating a personalized AI agent that has memory built in model weights. This way of memorizing requires some amount of computing and time, but will likely to equip agent with new skills and personalized character. In 6G AI agent services, all three methods can be utilized. In the following paragraphs, we will look closer at specific scenarios and how they require 6G network to manage its model and memory.

Fully computing offloaded AI agents: In scenarios of 6G network supporting wearable AI agent, the device is responsible for interacting with human while the network is responsible for running the agent and generating/storing memory as well as model fine-tuning. When user moves from one network service area to another, the model and memory should also transfer between 6G nodes. According to AI agent's service QoS, and its capability, multiple methods are available. First and foremost, 6G network could simply transfer agent model and memory as wearable device moves from 6G node 1 to node 2, but at cost of internal transmission link occupation. Second, when AI agent personalizing is convenient and fast enough, user equipment may help store memory and upload again to network when necessary. Thus, network may only cache the same base model, and let user upload its memory if it moves. Third, if memory file becomes much larger than agent model, and fine-tuning the model becomes an efficient way to digest those memory, instead of storing memories, wearable devices may choose to store its fine-tuned agent model and the network would only transfer additional memory not digested by agent model. Fig. 5 (1)-(3) illustrates the potential 3 scenarios.

6G Augmented AI agents: In scenarios of hand-held AI agents and robot/vehicles AI agents, user equipment is responsible for running its own AI agent locally, and network is responsible for facilitating local AI agent performance and providing additional information, such as short-term memory about the environment. Fig. 5 (4) gives an example of how 6G network could help fast moving autonomous vehicles to coordinate their routes and provide location-related context information as short-term memory. As the network has higher computing capability, it will serve local AI agent with its reflection process. Memory reflection allows an AI agents to independently summarize and infer more abstract, complex and high-level information. Fig. 5 (5) illustrates such scenario, where 6G provides necessary task model enrichment information to better complete planned tasks during the day and manages the AI agent memory and model at night.

Above scenario requires additional network capabilities on top of the currently envisioned design for 6G. First, higher network transmission speed is needed to facilitate frequently handle moving large models among network edges. Second, much more computing resources is expected at the network edge. Third, a set of functions intended for model and memory orchestration is desired.

Coordinated Agent-to-Agent Communication

In connected robotics and autonomous vehicle systems, one essential role of 6G agent is to facilitate application AI agent-to-agent communications whether through relaying or direct device-to-device (D2D) communication. One prominent feature of agent communication is that the source or destination are both application AI agents, and the data traffic packet size varies with the device computing capabilities, network conditions and the battery status. In some cases, it can be relatively small, examples are joint sensing or decision making information based on the local processing. While other cases when the local processing can not be done due to the lack of reliable data and device battery is not sufficient for local processing, the data traffic between agent to agent might be large. It is also worth mentioning that radio links are usually unstable and the

density and latency, reliability and transmission distance requirements are high with pervasive AI agents. This further increases the complexity of managing AI agent-oriented communication.

The communication and computing coordination among multiple AI agents is also vital. By monitoring the radio channel conditions, computing capabilities, battery/power supplying status and the location in real-time, 6G agent can help to dynamically form the computing and communication coordination cluster and topology. Multi-hop UE relays, direct D2D connection, and the connection and computing extension with moving vehicles or UAV BS with advanced computing capabilities can be flexibly configured with optimized resources

Another potential direction is leveraging the emerging semantic communication technology. As modern communication systems approach the limits of Shannon's theoretical transmission speed, semantic communication, likely the first method to achieve this breakthrough, focuses on encoding information in a more compact, compressed format. This format emphasizes the semantics of the original information over specific data, allowing it to be reassembled and understood at the receiving end. Semantic communication is suited to agent-to-agent communication, and large AI models plays a key role in the implementation of semantic communication, generating and deciphering multi-dimensional symbols for transmission. Before transmission, AI models can convert raw data into efficient symbolic representations, while at the receiving end, these symbols can be reassembled by AI models into the original data format.

For instance, in the transmission of images from an agent to another agent, a semantic communication framework based on diffusion models can synthesize and transmit image information retaining only the semantics at the sender's end. The receiver uses the diffusion model and spatially adaptive normalization to reconstruct semantically consistent information from this denoised semantic data.

Multi-Modal Data Transmission

A large part of AI agent consumed and generated contents are multi-modal, which generally includes AI model file, AI agent memory file (e.g. embedding vector database), sensor data, video, audio, text, etc. From their nature, we can see quite some different characteristics of data transmission in AI agent era.

- The synchronization of the multi-modal data is crucial for performing the accurate 3D object detection, ensuring the timely and precise decision making of the application AI agents and the collaboration of the multiple AI agents. Therefore, the future network needs to consider how to associate data flows related to multi-modal data. Policies containing the set of UEs and data flows, the expected QoS handling and coordination information can be provided by the AI agents. 6G network may leverage the obtained policies to coordinate the transmission of multiple UE's data flows for a multi-modal AI agent communication session. For example, the arrival delay difference threshold can be set for two data flows of one UE or multiple UE.
- The mutli-modal data generated in the AI agent era are envisaged to be more diverse. There is a need to enhance the QoS mechanism that allows the UE AI agent to proactively initiate or negotiate the desired QoS for multi-modal data so as to adaptively match the fluctuating radio conditions and dynamic device topology. This necessitates rethinking of: configuring and controlling entity of QoS flows, QoS profile, QoS monitoring mechanism.
- Multi-modal fusion. When transmitting multiple modal data flows at once, e.g., a video containing images and audio, finding equivalence among different modalities and effectively fusing them is crucial for optimizing communication efficiency and enhancing the quality of transmitted data. Using goal-oriented encoders, common and key features from various modalities, focusing on those most relevant to achieving a specific task or goal, can be extracted. While some modalities are particularly good at certain tasks, other modalities can be abstracted as supporting features to compliment missing information.

6G AI Agent Demonstration

In this section, we developed a proof of concept prototype for proposed 6G agent, as shown in Fig. 6, demonstrating the 6G Agent and Vehicle AI Agent collaboration use case. At the core of this prototype 6G agent is a state-of-the-art LLM, specifically GPT-4 turbo, which acts as the reasoning engine. LangChain framework [15] was used to design the peripherals of the AI agent, including a vector database for long-term memory and three network tools to handle requests from other application AI agents. RAG with FAISS was implemented to extract the most relevant information from vector database for requesting AI agents. The prototype was tested in a controlled environment based on code simulation. The following scenarios highlight the capabilities of the 6G AI agent.

In the first scenario, the 6G AI agent demonstrates its ability to provide real-time road status updates as context information to assist vehicle AI agent for safer autonomous driving. The 6G agent processes information collected from the all sorts of sources: sensors, cameras, vehicles and phones etc., established a road status vector databases, and abstract the road information that concerns requesting vehicle the most.

The second scenario focuses on the 6G agent's capability to configure network and satisfy customized network requests from application AI agents, such as the need for uninterrupted uplink video transmission, which critical for monitoring auto-driving. The vehicle AI agent requests assurance for an target uplink transmission speed. By forecasting the signal-to-noise ratio (SNR) along the vehicle route using "Link Predict" tool and preserving adequate network resources for the vehicle using the "SLA Assure" tool, the 6G AI agent guarantees the required transmission speed.

In the third scenario, the 6G AI agent addresses the need for dedicated D2D communication when overtaking another vehicle. The agent evaluate the required bandwidth and successfully allocates a private spectrum range,

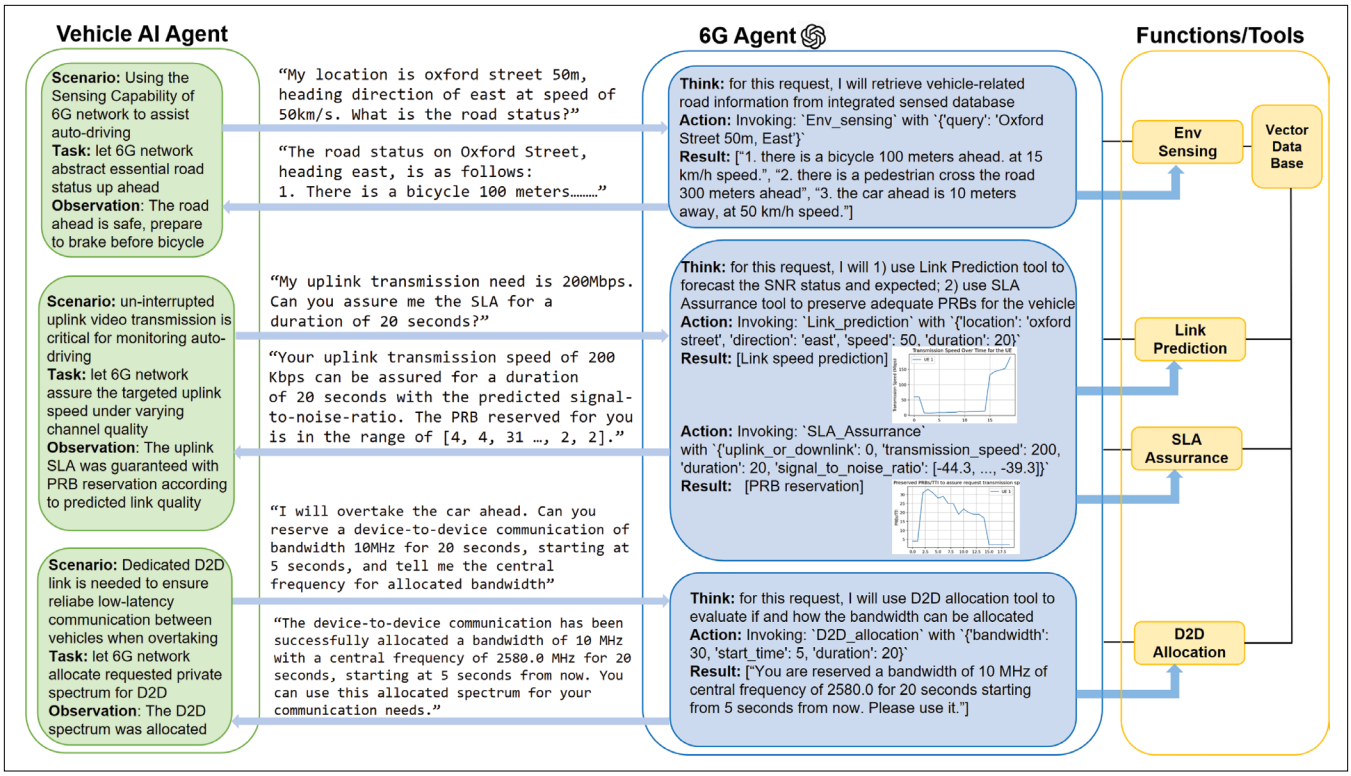

FIGURE 6. Demonstration of 6G Agent and Vehicle AI Agent collaboration.

ensuring prompt and reliable communication between vehicles.

These scenarios demonstrate the 6G AI agent's ability to use advanced tools and databases to fulfill complex tasks in real-time, showcasing the practical benefits of integrating AI agents with 6G networks. The demonstration highlights how AI agents can significantly enhance network performance, reliability, and user experience in various applications, paving the way for more sophisticated and intelligent 6G network services.

Challenges and Research Directions

Energy Saving and Edge Deployment

One of the most pressing issues of deploying LLM-based AI agents in 6G network is the high energy consumption associated with LLMs. Recent researches and engineering advancements have shown promising solutions to the high energy consumption of LLM inference. First is the constantly improving capability of small-scale LLMs. The most recent example is LLAMA-3 8B model, which was released by Meta in April 2024, outperforms its much larger predecessors such as LLAMA-2 70B, while consuming only a fraction of energy. Second, the Mixture of Experts (MoE) LLM architecture dynamically selects and activates only a subset of model components for each inference input, achieving nearly the same performance as the models using whole parameters while reducing much less inference cost and energy consumption. Lastly, new collaborative inference schemes and multi-exit models can help reduce processing time and cost by allowing intermediate results to be used when they meet the necessary accuracy thresholds. More researches into cloud-edge-end collaborations are expected to further solve the energy consumption and edge deployment challenge.

Security and Privacy Protection

AI agent is still a relatively new topic, and people are not fully understanding the full potential of AI agent, not to say its potential security risks. Here, we analysis some of the security risks that could happen in AI agent in 6G era.

- Data leakage: Data leakage in a 6G serviced AI agent presents a significant concern, primarily due to the vast amounts of sensitive data these agents generate and process, and the complex interactions with the network. The risk involves inadvertent exposure of confidential information, either through external breaches or internal vulnerabilities. To counter this, more researches into privacy-preserving techniques like data encryption and anonymization are expected to safeguard against potential data leakage.
- AI agent misbehaviour: If an AI agent misbehaves due to internal (self-evolution went wrong), or external reasons (agent kidnapped), it would pose significant risks due to its high-level access and control over network functionalities. Such a scenario could lead to unauthorized data access, network disruption, misuse of resources or even hazard situations. It is crucial to be very cautious about giving AI agent autonomy without human oversight, and giving AI agents recursive self-improvement abilities such as updates its own code and explore new environments. Therefore, robust security measures including continuous monitoring via Digital Twin, anomaly detection, and fail-safe protocols are

essential. Additionally, implementing stringent authentication and authorization processes can prevent unauthorized control or tampering with the AI agent.

AI Agent Tailored Communication

Currently LLMs are developed by various companies, and the way to interact with them are mainly natural language. However, when it comes to employing AI agents to realize robot-to-robot communication, human language shows its drawback: proportion of valuable key information is small, and lack of clarity. To ensure interoperability and efficiency, a common language allows different AI agents, potentially from diverse platforms and developers, to understand and exchange information seamlessly is essential. This uniformity is key in coordinating actions, sharing insights, and making collective decisions, especially in complex systems where multiple agents interact. Coding, compression and transmission of such language become important. 6G system is expected to meet the service requirements to enable communication between robots using media formats offering at least a better coding, compression and transmission efficiency than the that already specified for human consumption.

Conclusion

This article laid out a bold new vision for integration of AI agents within the 6G network, focusing on how 6G technologies enhance AI agent functionalities in various use case scenarios. We have proposed an AI agent-driven 6G network architecture, composed of service exposure layer, agent layer, function layer, and infrastructure layer. The paper also explores enabling technologies such as agent collaboration, model and memory management, agent-to-agent communication and multi-modal data transmission. Finally, it discussed the energy consumption and security challenges, and AI agent tailored communication as future research directions in this new 6G network, setting the stage for future developments in this rapidly evolving field.

References

[1] L. Wang et al., "A survey on large language model based autonomous agents," 2023, *arXiv:2308.11432*.
[2] F. Jiang et al., "Large language model enhanced multi-agent systems for 6G communications," 2023, *arXiv:2312.07850*.
[3] Y. Chen et al., "NetGPT: A native-AI network architecture beyond provisioning personalized generative services," 2023, *arXiv:2307.06148*.
[4] M. Xu et al., "Unleashing the power of edge-cloud generative AI in mobile networks: A survey of AIGC services," 2023, *arXiv:2303.16129*.
[5] Y. Huang et al., "Large language models for networking: Applications, enabling techniques, and challenges," 2023, *arXiv:2311.17474*.
[6] *Framework and Overall Objectives of the Future Development of IMT for 2030 and Beyond*, document Recommendation ITU-R IMT-2030, 2022.
[7] Q. Sun et al., "Intelligent ran automation for 5G and beyond," *IEEE Wireless Commun.*, vol. 31, no. 1, pp. 94–102, Feb. 2024.
[8] Y. Huang et al., "Validation of current O-RAN technologies and insights on the future evolution," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 2, pp. 487–505, Feb. 2024.
[9] V.-L. Nguyen et al., "Towards the age of intelligent vehicular networks for connected and autonomous vehicles in 6G," *IEEE Netw.*, vol. 37, no. 3, pp. 44–51, Jun. 2023.
[10] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, Oct. 2019.
[11] *Study on AI/ML Model Transfer Phase 2*, document TR 22.876 R19, 3GPP, 2023.
[12] *Study on Supporting Tactile and Multi-Modality*, document TR 22.847 R18, 2022.
[13] *Study on Network of Service Robots With Ambient Intelligence*, document TR 22.847 R19, 2023.
[14] W. Tong et al., "Ten issues of NetGPT," 2023, *arXiv:2311.13106*.
[15] O. Topsakal and T. C. Akinci, "Creating large language model applications utilizing LangChain: A primer on developing LLM apps fast," in *Proc. Int. Conf. Appl. Eng. Natural Sci.*, 2023, vol. 1, no. 1, pp. 1050–1056.

Biographies

Ziqi Chen received the B.E. degree in electric and communication systems from The Australian National University in 2015 and the Ph.D. degree in electrical engineering from the University of New South Wales in 2020. He is currently with the China Mobile Research Institute. His research interests include 5G networks, O-RAN, network automation, and generative AI. He is serving as the Co-Chair for the O-RAN Working Group 1 UCTG.

Qi Sun received the Ph.D. degree in information and communication engineering from the Beijing University of Posts and Telecommunications. She is currently a Senior Researcher with the China Mobile Research Institute. She has been working on the key technology and standardization of 5G radio access network, wireless big data and machine learning enabled 5G and future network architecture, and protocol and algorithm design. She has filed more than 20 patents, published more than 30 SCI/EI articles, and won the award of ESI top 1% highly cited article. She is the Chair of O-RAN Alliance Working Group 2 and had been served as the WG2 Chair of ITUT Focus Group on machine learning for 5G and future networks.

Nan Li (linan@chinamobile.com) received the master's degree from the Beijing University of Posts and Telecommunications in 2007, where he is currently pursuing the Ph.D. degree. In 2007, he joined the China Mobile Research Institute and has long been engaged in 4G/5G technology research, O-RAN network architecture, and protocol design. He has published more than ten articles in core journals and conferences, such as IEEE Wireless Communications and filed more than 300 patents.

Xiang Li joined Datang Telecom Company Ltd. in 2006, China Potevio in 2011, and the China Mobile Research Institute in 2016. He is currently a Project Researcher with the Department of Wireless and Terminal Technology, China Mobile Research Institute. His current research interests include integration of communication and computing on the RAN side and computing collaboration between mobile devices and RAN. He is the Co-Chair of O-RAN WG10.

Yang Wang received the B.S. and Ph.D. degrees from the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China, in 2018 and 2023, respectively. She is currently with the China Mobile Research Institute. Her current research interests include wireless sensing, statistical signal processing, and network optimization.

Chih-Lin I (Fellow, IEEE) is currently the CMCC Chief Scientist of Wireless Technologies. She has published over 200 papers in scientific journals, book chapters, and conferences, and holds over 100 patents. She is the co-author of the book *Green and Software-Defined Wireless Networks From Theory to Practice* (Cambridge University Press, 2019). She has also co-edited two books: *Ultra-Dense Networks: Principles and Applications* (Cambridge University Press, 2020) and *5G Networks: Fundamental Requirements, Enabling Technologies, and Operations Management* (Wiley-IEEE Press, 2018). Her current research interests center around ICDT Deep Convergence: "From Green & Soft to Open & Smart." She is a fellow of WWRF. She has won the 2005 IEEE ComSoc Stephen Rice Prize, the 2018 IEEE ComSoc Fred W. Ellersick Prize, the 7th IEEE Asia-Pacific Outstanding Paper Award, and the 2015 IEEE Industrial Innovation Award for Leadership and Innovation in Next-Generation Cellular Wireless Networks. She is the Chair of the O-RAN Technical Steering Committee and an O-RAN Executive Committee Member. She is the Chair of the FuTURE 5G/6G SIG and the Wireless AI Alliance (WAIA) Executive Committee; an Executive Board Member of GreenTouch; a Network Operator Council Founding Member of ETSI NFV; a Steering Board Member and the Vice Chair of WWRF; a Steering Committee Member and the Publication Chair of the IEEE 5G and Future Networks Initiatives; the Founding Chair of the IEEE WCNC Steering Committee; the Director of the IEEE ComSoc Meetings and Conferences Board; a member of IEEE ComSoc SDB, SPC, and CSCN-SC; and a Scientific Advisory Board Member of Singapore NRF.