Integrated AI and Communications: A Two-Way Catalysis Toward 6G and Beyond

Xiang Cheng, Jianan Zhang, Ning Ding, Nan Li, Yong Li, Tailin Wu, Wei Xu, Jun Zhang, Qi Sun

***Abstract*—Artificial intelligence (AI) and wireless communications are catalyzing each other's advancement as we approach 6G networks and beyond. This article presents a perspective on the two-way interplay between AI and communications—how AI techniques are revolutionizing communication network design (AI4Comm) and how emerging communication technologies are enabling and accelerating AI (Comm4AI). We discuss recent advances and outline the challenges in realizing an AI-native wireless ecosystem and propose a roadmap for integrating AI and communications, offering insights into a future where wireless communications and AI evolve together.**

***Keywords*—6G, artificial intelligence, AI-native networks, wireless communications, foundation models**

I. Introduction

As next-generation wireless networks will be enriched with pervasive multi-modal sensing and artificial intelligence (AI)-driven services, integrated sensing and communications (ISAC) and integrated AI and communications emerge as two key technological paradigms. To achieve AI-native intelligence integration of communication and multi-modal sensing, synesthesia of machines (SoM)[1] is a novel paradigm for future intelligent wireless networks. In particular, SoM-empowered next-generation wireless networks are envisioned to be AI-native, with AI deeply embedded in network design and operation. Meanwhile, the demands of large-scale AI services (e.g., foundation models and AI agents) are reshaping the design of communications and network system architectures. This two-way catalysis between AI and communications is poised to be a defining feature of 6G and beyond. In essence, AI will empower more efficient, adaptive, and autonomous communication networks (AI4Comm), while advances in communications and networks will support and accelerate AI model training, inference, and agent collaboration (Comm4AI).

Recent developments underscore this synergy. On the one hand, AI4Comm has yielded powerful data-driven solutions to longstanding wireless challenges. For example, deep learning and foundation models can predict and optimize wireless channels, protocols, and radio resource allocation beyond what classical approaches achieve. On the other hand, Comm4AI entails rethinking network design to serve AI workloads: distributing massive models across edge and cloud, minimizing latency for real-time AI inference, and reliably connecting a proliferation of intelligent agents. The convergence of these trends suggests that 6G networks will not only incorporate AI, but will be designed with AI as a service (AIaaS) as a key application scenario.

In this article, we first outline the dual paradigms of AI4Comm and Comm4AI, setting the stage for detailed highlights from the AI4Comm4AI seminar. Seminar speakers, listed alphabetically, include Xiang Cheng (Peking University), Ning Ding (Xi'an Jiaotong University), Nan Li (China Mobile Research Institute), Yong Li (Tsinghua University), Jing Liang (Huawei), Tailin Wu (Westlake University), Wei Xu (Southeast University), and Jun Zhang (The Hong Kong University of Science and Technology). We synthesize advances across both threads and connect them through the lens of SoM, the AI-native integration of multi-modal sensing

Manuscript received Sep. 16, 2025; revised Sep. 24, 2025; accepted Sep. 25, 2025. This work was supported in part by the National Natural Science Foundation of China under Grants 62341101, 62125101, and 62301011, in part by Beijing Natural Science Foundation under Grant L257016, and in part by the New Cornerstone Science Foundation through the Xplorer Prize. The associate editor coordinating the review of this paper and approving it for publication was L. Q. Fu.

X. Cheng, J. N. Zhang. State Key Laboratory of Photonics and Communications, School of Electronics, Peking University, Beijing 100871, China (e-mail: xiangcheng@pku.edu.cn; zhangjianan@pku.edu.cn).

N. Ding. Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: ding.ning@xjtu.edu.cn).

N. Li, Q. Sun. China Mobile Research Institute, Beijing 100053, China (e-mail: linan@chinamobile.com; sunqiyjy@chinamobile.com).

Y. Li. Department of Electronic Engineering, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: liyong07@tsinghua.edu.cn).

T. L. Wu. Department of Artificial Intelligence, Westlake University, Hangzhou 310030, China (e-mail: wutailin@westlake.edu.cn).

W. Xu. National Mobile Communications Research Laboratory, Purple Mountain Laboratories, Southeast University, Nanjing 210096, China (e-mail: wxu@seu.edu.cn).

J. Zhang. Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong 999077, China (e-mail: eejzhang@ust.hk).

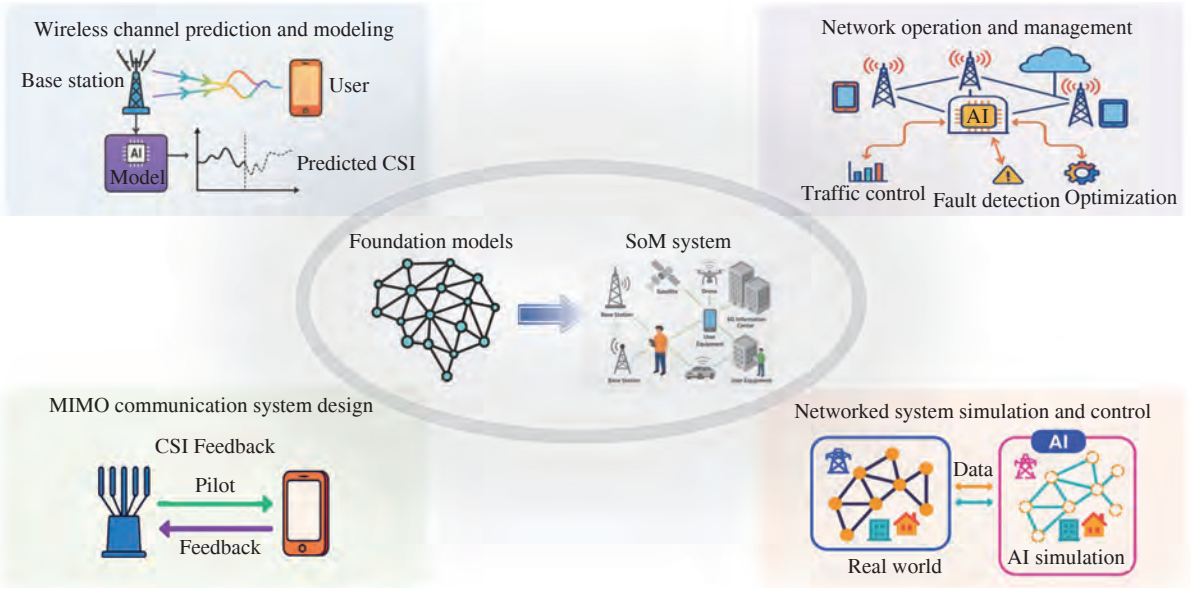


Fig. 1 AI for communications

and communications that underpins intelligent 6G services. We highlight progress in AI4Comm (foundation models for wireless communications and multiple-input multiple-output (MIMO) design, network operation and management, and networked system simulation and control) and Comm4AI (6G architectures to host large models and agents, wireless networks for edge learning, and core networks for large model training). We then list the challenges spanning data, model, computing and infrastructure, trust, safety and security, and standardization and integration that must be addressed to realize AI-native intelligence. Accordingly, we propose a three-phase roadmap that evolves from AI-enhanced 5G to an AI-integrated 6G architecture and ultimately to AI-native 6G autonomously orchestrated by networked AI agents. We conclude with an outlook on the co-evolution of AI and communications.

II. AI4Comm: AI for Communications

AI4Comm refers to leveraging AI techniques to solve problems in communications and networks and advance their system designs. This spans a wide range of applications: using AI for channel estimation and prediction, optimizing antenna beamforming with neural networks, managing network traffic with AI, and designing protocols that learn and adapt. Early applications of AI in wireless communications focused on training deep neural networks for specific tasks. Now, attention is shifting to foundation models and cross-domain AI that can handle multiple tasks or even entire layers of the network stack. Ref. [2] presents the first systematic investigation and framework for designing foundation model-enhanced SoM systems. It emphasizes the significant potential of foundation models to address key challenges within SoM, highlighting their promising capability to transform future communication and network infrastructure design paradigms. The generalization and reasoning abilities of such models make them attractive for complex communication scenarios that are hard to explicitly model. These aspects are depicted in Fig. 1 and will be elaborated in the remainder of this section.

A. Foundation Models for Wireless Channel Prediction and Modeling

One prominent example of AI4Comm is the adaptation of LLMs for wireless channel prediction. A transformer-based LLM can be fine-tuned to predict wireless channel state information (CSI) sequences by dedicated design to bridge the gap between channel data and the feature space of the LLM[3]. The model, termed as large language model for channel prediction (LLM4CP), leverages a pre-trained LLM as a powerful predictor of future CSI based on past observations[3]. By transferring knowledge from vast text-based pre-training to the wireless domain, LLM4CP captured complex temporal patterns in channel fading that conventional methods (e.g., Kalman filters or simple recurrent models) struggle with. LLM4CP achieved superior accuracy in predicting channel variations (e.g., due to user mobility) and reduced prediction error significantly compared to baselines. This work illustrates the promise of fine-tuning state-of-the-art AI models for communication-specific tasks.

A wireless foundation model WiFo is designed specifically for wireless communication[4]. WiFo uses a transformer-based masked autoencoder architecture aimed at providing a universal solution for many channel-related tasks. During self-supervised pre-training, portions of the CSI data across time, frequency, and antenna domains are masked, and the

model learns to reconstruct the missing elements. Through this process, WiFo acquires a robust internal representation of wireless channels and can handle tasks such as time-series extrapolation and frequency-domain interpolation within a single model. The ability to deal with arbitrary patterns of missing data implies that one foundation model can replace many task-specific models, greatly simplifying network management. In evaluations, WiFo approached or exceeded the accuracy of specialized deep networks on various prediction tasks, while being more flexible. These results mark a shift from specific AI solutions toward broad AI platforms for communications.

B. Foundation Models for MIMO Communication System Design

Physical layer tasks such as channel estimation, beamforming and detection interact with the same radio channel, yet today's practice trains a separate neural network for each task. To remove this redundancy, CSI Foundation Models[5] furnish a lightweight, physics-aware prior for all transceiver modules. It has a single generative prior that learns the score function of wide-band MIMO channels from limited clean CSI or even raw pilot data. By injecting labels such as user location, link state or weather into the network, one model can synthesize channels for many scenarios without retraining, thereby improving sample efficiency and deployment simplicity. When moving to a new cell, parameter-efficient fine-tuning updates only a small subset of weights, preserving latency budgets while coping with local propagation idiosyncrasies. Once trained, the score network can be plugged into diverse downstream modules as a universal prior, dramatically cutting model count, memory footprint and life-cycle cost.

In addition, AI can reduce feedback overhead and boost channel recovery in frequency division duplexing (FDD) massive MIMO. DrCsiNet[6] is a variational auto-encoder framework that disentangles downlink CSI into two parts: an exclusive representation that is unique to the downlink and a shared representation that is implicitly present in both uplink and downlink channels. An encoder at the user equipment extracts and feeds back only the exclusive codeword, while two networks at the base station derive the shared component directly from uplink pilots, eliminating redundant transmission. A representation recovery decoder then fuses both latents to reconstruct full downlink CSI. Additionally, researchers have integrated the intrinsic physical characteristics of CSI, such as spatial and temporal correlations, as well as amplitude and phase features. By leveraging this expert prior information in AI-based CSI compression feedback methods, feedback overhead and model complexity have been effectively reduced. Thus, AI replaces rigid codebooks with adaptive, reciprocity-aware compression that scales to larger MIMO arrays.

C. Foundation Models for Network Operation and Management

LLM is able to coordinate multiple task-specific algorithms already embedded in production networks. An LLM interprets operator intent in natural language, decomposes it into executable plans, invokes specialized monitoring or repair tools, records outcomes in long-term memory and iteratively refines actions until key performance indicators are met. This enables zero-touch deployment, fault localization and self healing across radio access network (RAN) and core domains[7]. Ultimately, the architecture evolves from many siloed models toward a multi-scene, multi-task unified model that delivers intent driven, closed-loop network operation and paves the way for AI-native 6G services.

Rather than treating machine learning modules as add-ons, a dedicated AI functional framework can be native to 6G system design. The framework introduces an AI coordination and optimization function that orchestrates training, inference and performance monitoring across distributed base station and core nodes, thereby enabling closed-loop, data-driven optimization of mobility, load balancing, energy saving and other RAN procedures. A suite of logical functions can operationalize this framework. The intelligent coordination controller issues multi-dimensional resource policies; the computing control/service functions and data control/service functions manage edge compute and data pipelines; while a model management function governs life-cycle tasks such as versioning, retraining and security. These components support hierarchical and distributed intelligence, allowing lightweight inference to remain at latency-critical layer-1 and layer-2 points while heavier learning tasks are pooled at edge or cloud aggregators[7-8].

Motivated by the shift from specialized, task-centric algorithms to general-purpose agents capable of reasoning, multi-modal perception and autonomous decision-making, WirelessAgent[9] is a unified agent architecture that harnesses LLMs to automate complex 6G network operations. The design principles include interaction with users, environment and peer agents; autonomy to act without step-by-step human supervision; and self-improvement via continual learning from feedback and new data. WirelessAgent realizes these principles through four tightly-coupled modules. Perception converts natural-language instructions and raw wireless measurements (e.g., CSI, vision, location) into structured internal representations, relying on LLM text understanding plus lightweight translators for non-text modalities. Memory persistently stores observations and past actions, enabling rapid recall of similar situations and reducing repetitive mistakes. Planning decomposes a high-level goal into sub-tasks using LLM reasoning, augments knowledge via retrieval-augmented generation, and performs reflection be-

fore and after acting to refine future behaviors. Action expresses decisions through natural-language responses and tool calls. A case study shows that WirelessAgent achieves lower resource occupancy, higher slice capacity and adaptive reallocation.

D. Foundation Models for Networked System Simulation and Control

Network Digital Twins create high-fidelity virtual replicas of mobile networks so that operators can test various policies without risking live traffic. Generative AI and foundation models transform every stage of the network digital twin life-cycle, yielding faster, more accurate and more adaptive simulations[10]. For data processing and network monitoring, generative models fill spatio-temporal gaps, impute corrupted measurements and flag anomalies, which raises data fidelity without expensive additional probes. For digital replication and simulation, generative models learn the joint distribution of heterogeneous network variables, including traffic, radio maps, user mobility, and sample synthetic trajectories at orders-of-magnitude higher speed while preserving realistic correlations. For optimizer design and training, high-throughput AI-driven simulators supply virtually unlimited training data to reinforcement learning (RL) agents that search spectrum, power or tilt settings. Diffusion-model-assisted RL further expands the exploration space, improving sample efficiency and solution quality. For Sim2Real transition and control, generative AI continuously refines the digital twin with real telemetry and mitigates the reality gap, allowing policies learned in simulation to be safely deployed and updated online.

Both simulation and control can be cast as probabilistic generation tasks. Diffusion Physical systems Control (DiffPhyCon) unifies simulation and control for dynamic physical systems[11]. It is first pre-trained on large, heterogeneous trajectories and learns a joint generative distribution over system states and control signals. At deployment, objectives, safety limits and other constraints are injected as extra energies whose gradients are added to the learned score, so one pretrained model can be steered toward new tasks, geometries or operating regimes without retraining. The closed-loop successor CL-DiffPhyCon[12] decouples the denoising schedule along the physical horizon: early actions are generated with lower noise, executed on the system, and the measured state feeds back into the trajectory. This asynchronous scheme preserves performance while cutting sampling cost by up to an order of magnitude, enabling real-time, robust control of complex systems.

In summary, foundation models have been applied to a wide range of wireless communication network tasks, delivering gains in prediction accuracy and system efficiency. Foundation models exhibit strong generalization when they capture the underlying wireless patterns and transfer knowledge across domains. For example, a self-supervised foundation model can learn a robust internal representation of channel dynamics that supports multiple downstream tasks. A large model pre-trained on broad data can be fine-tuned to recognize complex wireless signal patterns that specialized models miss. At scale, such models exhibit emergent reasoning and the capacity to abstract common principles (e.g., radio propagation regularities), which underpins cross-task generalization.

III. Comm4AI: Communications for AI

Comm4AI refers to innovations in communication networks specifically aimed at supporting AI applications and workflows. Task-oriented elastic networks play a crucial role in supporting complex SoM processing, as they leverage heterogeneous resources and dynamically adapt to resource availability. As AI models grow in scale and complexity, and as AI services become more pervasive, future networks must meet new requirements in data delivery, latency, and reliability to enable these AI systems. These aspects are illustrated in Fig. 2 and will be introduced in this section.

A. 6G for Artificial General Intelligence

6G is envisioned as an AI-native network that tightly integrates communication with computing and sensing. Rather than acting solely as a data pipe, 6G forms a unified communication, sensing, and AI platform. A representative example is the AI-integrated radio access network, which embeds core agent capabilities, such as sensing, cognition, decision-making, and action, directly into the air interface. The resulting network functions as a neural center that links vast numbers of intelligent agents, supporting their learning, training, and inference. By distributing AI processing throughout the infrastructure, 6G supplies both the computational power and the ultra-low latency connectivity required to host advanced AI models across edge and cloud resources. The 6G core network adopts an agent-based, task-oriented architecture. Every network element may host an intelligent agent, allowing the infrastructure to self-optimize around application intents. Customized network slices are instantiated on demand and torn down when no longer needed. Functionally, the 6G core operates as a distributed computing platform that spans devices, edge, and cloud. Edge devices run lightweight AI for local sensing and fast inference. Cloud and core house large foundation models. 6G links orchestrate the two in real time, creating seamless device-pipe-cloud synergy. This agent-based, distributed design lets the network coordinate myriad AI agents with minimal latency, making 6G an essential enabler of artificial general intelligence (AGI) by providing the intelligent, adaptive fabric for collective, general-purpose intelligence[13].

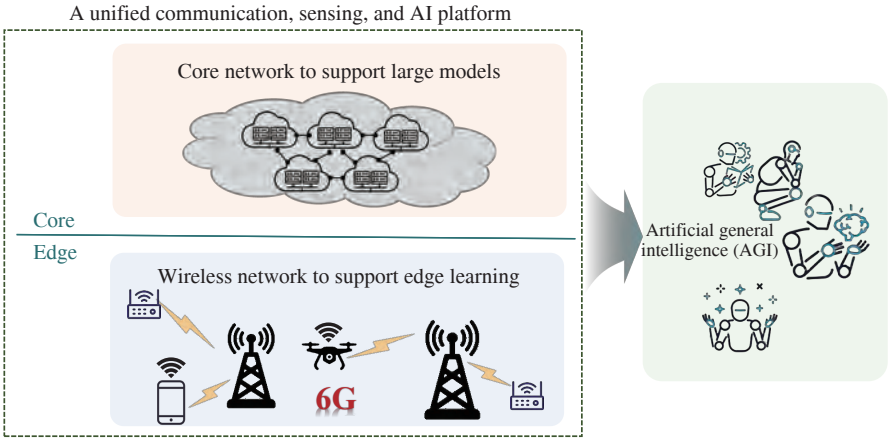


Fig. 2 Communications for AI

Data quality in the network heavily influences the success of AGI algorithms deployed on it. A data quality assessment framework tailored for wireless air-interface evaluates wireless datasets (e.g., channel state information logs, signal metrics, etc.) for similarity and diversity[14]. By quantifying these aspects, the network can screen and improve its data before using it for AI model training or inference. By measuring similarity and diversity, one can select or even generate better training data (e.g., channel state samples) that yield more accurate compression or prediction models. In summary, data quality management is important and the infrastructure will need mechanisms to evaluate incoming data streams and curate datasets so that the AI brain of the network learns from high-fidelity, representative information.

B. Wireless Network to Support Edge Learning

Wireless network design plays a pivotal role in supporting edge learning by ensuring that communication systems effectively accommodate the challenges of distributed learning tasks. A key aspect is optimizing resource allocation to balance communication and learning tasks. Since edge learning relies on frequent data exchanges and parameter updates between devices and central aggregators, communication resources such as bandwidth, power control, and scheduling must be carefully managed to reduce latency and overhead. Furthermore, edge learning systems often face constraints on wireless channels, such as fading and interference, which can degrade learning performance. In this regard, comparative studies of digital versus analog transmission schemes for wireless federated learning (FL) have provided new insights into the tradeoffs between communication and learning performance. Advanced methods such as beamforming and dynamic scheduling are necessary to enhance the spectral efficiency of the network and reduce communication delays.

To address these challenges, edge learning networks require joint optimization frameworks that integrate machine learning algorithms with network design, allowing efficient computation offloading and task prioritization across devices. Intelligent edge devices, base stations, and the wireless medium collaboratively improve learning and inference performance. Instead of optimizing communication for maximum throughput alone, networks will be engineered to maximize AI task performance. This includes jointly designing signaling, processing algorithms, and wireless resources to meet AI-driven metrics, such as model accuracy or decision latency, alongside traditional metrics[15]. In parallel, energy-efficient edge inference frameworks for integrated sensing, communication, and computation (ISCC) networks have been proposed, where techniques such as split inference, model pruning, and feature quantization are jointly optimized to reduce energy consumption under stringent latency and accuracy requirements. Such joint optimization also applies to multi-agent embodied AI which requires real-time learning and coordination[16].

C. Core Network to Support Large Models

Training cutting-edge AI (such as billion-parameter neural networks or federated learning across many devices) often involves splitting computation across multiple servers or edge devices. This distributed training demands intensive data exchange between nodes (for gradient updates, model parameters, etc.), which can become a bottleneck. Co-optimizing computation and communication improves resource utilization. For example, scheduling the forward and backward passes of neural network training in a bidirectional, overlapping manner can improve the utilization of network links. Beyond static pipelines, adaptive resource allocation in distributed AI further improves resource utilization. Farseer predicts an increasing training data to model size ratio as compute budget increases, which aligns with the actual training configurations of recent state-of-the-art LLMs, and facilitates more efficient evaluation of compute allocation[17].

In summary, Comm4AI research recognizes that future AI workloads, including training massive models, serving distributed AI applications, or running AI-driven control loops, will place demands on networks that go beyond what current architectures can handle. By innovating in areas like distributed training protocols, network scheduling for AI traffic, and AI-native architectural principles, we can ensure that networks serve as a powerful platform for AI evolution.

IV. Challenges

Although the synergy between AI and communications presents significant potential, several critical challenges remain that must be addressed to achieve this envisioned future. We discuss some of the key open issues at the intersection of AI and communications.

A. Data

AI4Comm solutions typically require extensive datasets, including channel samples and multi-modal sensor inputs, for effective training. In particular, wireless foundation model-based approaches demand significantly greater data diversity and volume compared to conventional task-specific models. However, obtaining high-quality, representative datasets for wireless environments is challenging due to the cost of extensive measurements and the difficulty of integrating multiple simulation software. Specifically, precise spatio-temporal alignment across multiple modalities imposes additional challenges for constructing communication and multi-modal sensing datasets[18] used in SoM research.

B. Model

Although foundation models promise to revolutionize the design paradigms of AI4Comm and Comm4AI, they encounter significant challenges related to multi-modality and task heterogeneity. On the one hand, to effectively support future SoM systems, foundation models must simultaneously process large-scale, diverse wireless and multi-modal sensing data, including CSI, RGB images, depth maps, mmWave (millimeter wave) radar and light detection and ranging (LiDAR) signals. The variations in dimensionality and distribution across these data types present considerable challenges for joint processing. On the other hand, foundation models are also required to concurrently address a broad spectrum of AI4Comm and Comm4AI tasks. Tasks differ in their input-output formats, objectives, mechanisms, and operational timescales. Consequently, they demand distinct tokenization and encoding schemes, input adapters, output heads, and loss functions, while imposing heterogeneous latency and robustness requirements. Such factors collectively complicate unified modeling and end-to-end training.

Moreover, large AI models are expensive to train and deploy. Their size raises questions about feasibility at network edge devices. Compressing models via pruning, quantization, or knowledge distillation without sacrificing performance will be crucial if we expect edge nodes or user devices to run these large models locally. There is also the risk of model drift over time as network conditions evolve; continual learning frameworks might be needed so that models can update themselves with new data. Furthermore, combining prior knowledge from communication tasks with model design can create a network that is not solely data-driven, thereby effectively reducing its complexity. For example, the weighted minimum mean square error (WMMSE) algorithm can be transformed into a neural network, and the self-information of CSI can be leveraged to enhance CSI compression and encoding.

C. Computing and Infrastructure

Realizing AI4Comm and Comm4AI at scale will require redesigning the network infrastructure. On the one hand, embedding AI into many parts of the network (edge inference, intelligent radio units, core network optimizers) calls for pervasive computing resources. This may drive a need for deploying AI accelerators at base stations or aggregation points, increasing cost and energy consumption. Efficient hardware utilization, possibly through virtualization and sharing of compute across multiple AI tasks, will be important. On the other hand, supporting AI applications such as distributed training or augmented reality/virtual reality (AR/VR) requires new network services. These could include dedicated high-bandwidth links on demand, in-network caching of AI model parameters, or localized data centers for edge AI tasks. There is also a challenge of scalability—many AI4Comm approaches have been demonstrated in relatively controlled scenarios. It remains to be seen how they perform in large, heterogeneous networks with thousands of nodes. Scalability in terms of both algorithm complexity and orchestration (coordinating learning across many devices) is an open issue.

D. Trust, Safety, and Security

Introducing AI into the control loop of networks implies that we must trust these models to operate correctly under all conditions[19]. However, AI models may lack formal guarantees, and ensuring the reliability and safety of AI decisions is paramount. One approach is uncertainty quantification for model outputs. By producing calibrated predictive distributions or confidence sets, the network can attach risk-aware confidence to channel predictions, anomaly scores, and control actions. Decisions with high uncertainty can be gated to conservative fallbacks or human oversight, and constraint sets can be tightened adaptively when uncertainty spikes. Adversarial attacks are another concern: an attacker might spoof inputs (such as falsifying sensor data or CSI feedback) to

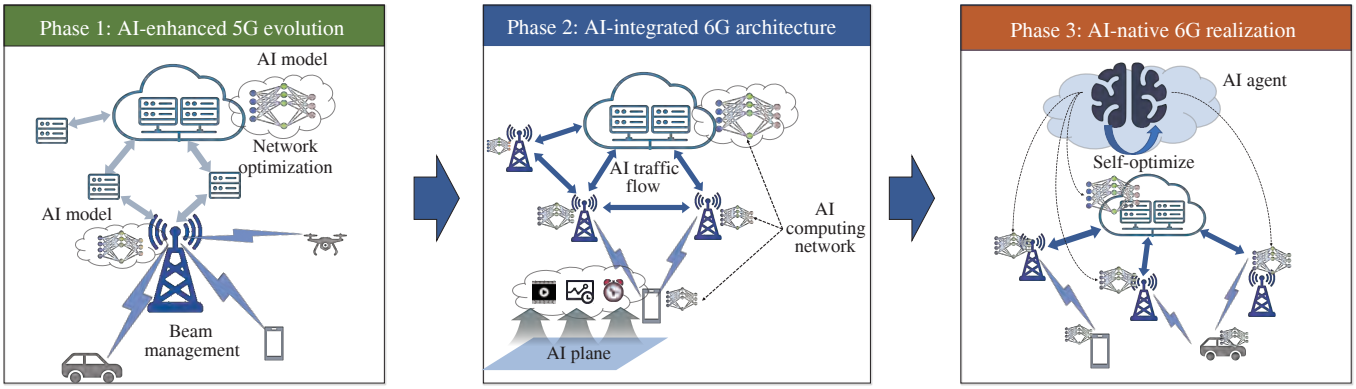


Fig. 3 Roadmap toward AI-native 6G systems

mislead an AI model controlling the network. Robustness against such attacks and secure data exchange for distributed AI need to be built in. AI algorithms could also inadvertently reinforce biases (e.g., unfair resource allocation) present in training data, leading to ethical and regulatory issues in network management. In wireless federated learning, recent works have explored reconfigurable intelligent surfaces (RIS)-based interference suppression to ensure unbiased over-the-air aggregation[20], as well as zero-trust Byzantine-resilient frameworks with adaptive clustering to defend against malicious devices, highlighting practical approaches toward robustness. Thus, alongside performance, aspects of fairness, accountability, and transparency of AI in communications must be researched.

E. Standardization and Integration

For AI to truly become native to 6G, standardization bodies (3GPP, ITU, etc.) will need to define interfaces and frameworks that integrate AI algorithms with traditional network functions. Communications systems are traditionally designed with rigorous specifications, whereas AI components are probabilistic and data-driven. Defining standards for how models and data are exchanged between network entities will be necessary to ensure interoperability. Additionally, the development cycle for AI models is different from that of network protocols; networks might need to accommodate frequent updates of AI models without disrupting service. The collaboration between the telecom industry and AI research community will be essential to address these integration challenges.

V. Roadmap Toward AI-Native 6G Systems

Drawing on the recent developments and the challenges outlined above, we sketch a high-level roadmap for achieving the two-way AI-communications integration in 6G. The path forward is framed in three phases, as shown in Fig. 3.

A. Phase 1: AI-Enhanced 5G Evolution (Short Term)

The focus of Phase 1 is on incremental enhancements: bringing AI into current-generation (5G/5G-Advanced) networks in carefully scoped ways. This includes standardizing network data analytics functions, applying machine learning for network optimization (e.g., traffic prediction for dynamic spectrum allocation), and developing edge platforms for hosting AI applications (such as AR/VR processing offload). Importantly, this phase involves extensive prototyping and field tests to build confidence in AI techniques. We expect to see "AI in the loop" for specific use cases, such as AI-assisted mmWave beam management or energy saving features, where learning-based algorithms run in parallel with legacy algorithms for safety. The near term also involves defining preliminary standards for AI in networks, for example, 3GPP's work on AI/ML model exchange, and addressing regulatory questions about AI decisions in telecom services.

B. Phase 2: AI-Integrated 6G Architecture (Medium Term)

As we move into the 6G era, AI is expected to be more deeply integrated into the network architecture. In this phase, we envision the emergence of a true AI plane in the network: a layer of functionality dedicated to learning and inference tasks that support both network operations and user-facing AI services. Concretely, 6G standards might include native support for model distribution (how AI models are broadcast, updated, and stored at various network nodes) and federated learning (enabling on-device learning with privacy protection). The network infrastructure will likely incorporate distributed computing fabric, essentially turning base stations into micro datacenters, so that latency-sensitive AI tasks can be executed close to users. Networks will begin offering AI quality of ser-

vice (QoS) guarantees, such as ensuring a certain latency/jitter bound for AR communication, and perhaps isolating AI traffic flows to maintain reliability. Standardization efforts will solidify around formats for data and knowledge sharing between operators' networks and AI service providers.

C. Phase 3: AI-Native 6G Realization (Long Term)

The long-term vision is to have fully AI-native networks in operation. In this phase, many network functions could be autonomously managed by AI agents. For example, a 6G system might continuously self-optimize using reinforcement learning agents that tweak parameters (power, spectrum allocation, routing) in real time, with minimal human intervention. The network will expose application program interfaces (APIs) for third-party AI applications, making the communication fabric highly programmable. We might see scenarios such as on-demand deployment of an AI-driven network slice, for instance, a slice optimized for an incoming swarm of autonomous drones, configured largely by AI analyzing the situation. Reliability and safety measures for these AI systems will be entrenched—networks could have fail-safes that revert to conservative operation if an AI anomaly is detected. Explainability tools may be integrated so that if regulators or operators query a decision, the system can provide rationale. By this phase, the human role in direct network control will be supervisory; engineers will spend more time defining objectives and constraints for AI controllers, rather than tuning low-level algorithms. Achieving this level of autonomy will likely require surmounting significant technical and trust barriers, but if done successfully, it promises a network that is far more adaptive and efficient than today's static configurations.

Realizing a single foundation model that generalizes across the full spectrum of wireless communication network tasks remains a long-term objective. Multi-modal datasets measured and simulated across different frequency bands, mobility patterns, hardware platforms, and sensing modalities are essential for robust operation in diverse scenarios. Adaptation can be strengthened by attaching domain and task-specific adapters to a shared backbone and by employing mixture-of-experts architectures that route tasks to specialized submodules. Coupled with physics-aware objectives and continual learning to track distribution shifts, such models can serve as general-purpose controllers for communication network systems.

VI. Conclusion

The convergence of AI and wireless communications is poised to define the trajectory of 6G networks. AI4Comm and Comm4AI form a virtuous cycle: advanced AI algorithms unlock new levels of network performance and automation, while next-generation networks provide the high-speed, low-latency networks that AI systems need to reach their full potential. SoM enhances the AI-native integration of communication and multi-modal sensing and accelerates the co-evolution of AI and communications. We highlighted how foundation models are adapted to predict channels, simulate networked systems, and assist in network management. We also saw how communication principles are influencing AI system design via techniques such as communication-efficient distributed training.

Looking ahead, achieving an AI-native 6G systems will require sustained interdisciplinary collaboration. Wireless experts and AI researchers must collaborate to design models that respect communication constraints and networks that are flexible enough to host AI services. There are still significant challenges to overcome, ensuring the reliability, security, and fairness of AI-driven decisions in networks will be as important as achieving raw performance gains. The roadmap we outlined suggests a phased approach—experiment and standardize in the near term, integrate and optimize in the medium term, and finally deploy at scale with built-in learning and adaptation.

In conclusion, the two-way catalysis of AI and communications offers a compelling vision for the future. AI techniques will continue to revolutionize how networks are designed and operated, and advances in communication technology will, in turn, enable AI to pervade every aspect of our lives. The 6G era will likely witness this symbiotic relationship delivering networks that are more adaptive, efficient, and capable than ever before.

Conflict of interest statement. None declared.

References

[1] CHENG X, ZHANG H, ZHANG J, et al. Intelligent multi-modal sensing-communication integration: synesthesia of machines[J]. IEEE Communications Surveys & Tutorials, 2023, 26(1): 258-301.

[2] CHENG X, LIU B, LIU X, et al. Foundation model empowered synesthesia of machines (SoM): AI-native intelligent multi-modal sensing-communication integration[J]. arXiv preprint arXiv:2506.07647, 2025.

[3] LIU B, LIU X, GAO S, et al. LLM4CP: adapting large language models for channel prediction[J]. Journal of Communications and Information Networks, 2024, 9(2): 113-125.

[4] LIU B, GAO S, LIU X, et al. WiFo: wireless foundation model for channel prediction[J]. Science China Information Sciences, 2025, 68(6): 162302.

[5] YU W, HE H, SONG S, et al. AI and deep learning for THz ultra-massive MIMO: from model-driven approaches to foundation models[J]. arXiv preprint arXiv:2412.09839, 2024.

[6] XU W, WU J, JIN S, et al. Disentangled representation learning empowered CSI feedback using implicit channel reciprocity in FDD massive MIMO[J]. IEEE Transactions on Wireless Communications, 2024, 23(10): 15169-15184.

[7] LI N, WANG Y, SUN Q, et al. Rethinking RAN architecture for deep fusion of AI and communication in 6G[J]. IEEE Wireless Communications, 2025, (32): 3.

[8] SUN Q, LI N, CHIH-LIN I, et al. Intelligent RAN automation for 5G and beyond[J]. IEEE Wireless Communications, 2023, 31(1): 94-102.
[9] TONG J, GUO W, SHAO J, et al. WirelessAgent: large language model agents for intelligent wireless networks[J]. arXiv preprint arXiv:2505.01074, 2025.
[10] LI T, LONG Q, CHAI H, et al. Generative AI empowered network digital twins: architecture, technologies, and applications[J]. ACM Computing Surveys, 2025, 57(6): 1-43.
[11] WEI L, HU P, FENG R, et al. DiffPhyCon: a generative approach to control complex physical systems[C]//Proceedings of the Advances in Neural Information Processing Systems. Red Hook: Curran Associates Inc. 2024: 4090-4147.
[12] WEI L, FENG H, YANG Y, et al. CL-DiffPhyCon: closed-loop diffusion control of complex physical systems[C]//Proceedings of the International Conference on Learning Representations. United Kingdom: PMLR, 2025.
[13] TONG W, MA J, ZHU P, et al. AI: the bridge to 6G[J]. Huawei Tech, 2024: 1.
[14] TANG H, YANG L, ZHOU R, et al. Assessing air-interface dataset similarity and diversity for AI-enabled wireless communications[C]//Proceedings of the IEEE International Conference on Communications Workshops (ICC Workshops). Piscataway: IEEE Press, 2024: 1623-1628.
[15] XU W, YANG Z, NG D W K, et al. Edge learning for B5G networks with distributed signal processing: semantic communication, edge computing, and wireless sensing[J]. IEEE Journal of Selected Topics in Signal Processing, 2023, 17(1): 9-39.
[16] FENG Z, XUE R, YUAN L, et al. Multi-agent embodied AI: advances and future directions[J]. arXiv preprint arXiv:2505.05108, 2025.
[17] LI H, ZHENG W, WANG Q, et al. Farseer: a refined scaling law in large language models[J]. arXiv preprint arXiv:2506.10972, 2025.
[18] CHENG X, HUANG Z, YU Y, et al. SynthSoM: a synthetic intelligent multi-modal sensing-communication dataset for synesthesia of machines (SoM)[J]. Scientific Data, 2025, 12(1): 819.
[19] YANG Z, XU W, LIANG L, et al. On privacy, security, and trustworthiness in distributed wireless large AI models[J]. Science China Information Sciences, 2025, 68(7): 1-15.
[20] SHI W, YAO J, XU W, et al. Combating interference for over-the-air federated learning: a statistical approach via RIS[J]. IEEE Transactions on Signal Processing, 2025, 73: 936-953.

About the Authors

Xiang Cheng [corresponding author] received the joint Ph.D. degree from Heriot-Watt University and The University of Edinburgh, Edinburgh, UK, in 2009. He is currently a Boya Distinguished Professor with Peking University, Beijing, China. His research focuses on the in-depth integration of communication networks and artificial intelligence, including intelligent communication networks and connected intelligence, the subject on which he has published more than 280 journals and conference papers, 11 books, and holds 32 patents. He was a recipient of the IEEE Asia-Pacific Outstanding Young Researcher Award in 2015 and the Xplorer Prize in 2023. He was a co-recipient of the 2016 IEEE Journal on Selected Areas in Communications Best Paper Award: Leonard G. Abraham Prize and the 2021 IET Communications Best Paper Award: Premium Award. He has also received the Best Paper Awards at IEEE ITST'12, ICCC'13, ITSC'14, ICC'16, ICNC'17, GLOBECOM'18, ICCS'18, and ICC'19. He has been a Highly Cited Chinese Researcher since 2020. In 2021 and 2023, he was selected into two world scientist lists, including the World's Top 2% Scientists released by Stanford University and top computer science scientists released by Guide2Research. He has served as the symposium lead chair, the co-chair, and a member of the technical program committee for several international conferences. He led the establishment of four Chinese standards (including industry standards and group standards) and participated in the formulation of ten 3GPP international standards and two Chinese industry standards. He is currently a Subject Editor of IET Communications; an Associate Editor of IEEE Transactions on Wireless Communications, IEEE Transactions on Intelligent Transportation Systems, IEEE Wireless Communications Letters, and Journal of Communications and Information Networks. He was a Distinguished Lecturer of the IEEE Vehicular Technology Society and a Fellow of IEEE.

Jianan Zhang received the B.E. degree from Tsinghua University, Beijing, China, in 2012, and the M.S. and Ph.D. degrees from Massachusetts Institute of Technology, Cambridge, USA, in 2014 and 2018, respectively. He is currently an Assistant Professor with the School of Electronics, Peking University, Beijing, China. Prior to joining Peking University, he was a Senior Software Engineer with Google LLC, Sunnyvale, CA, USA, from 2018 to 2023. His research interests include network theory and networked intelligence, with focus on resource allocations and foundation models for network optimizations. His research has applications to distributed machine learning systems, networked autonomous systems, cyber physical systems, and data center networks. He is a co-recipient of the WiOpt 2024 Best Paper Award. He serves as an Area Editor of Computer Networks.

Ning Ding received the B.S. (2005) and M.S. (2008) degrees from Xi'an Jiaotong University, Xi'an, China, and the Ph.D. degree in 2012 from Keio University, Tokyo, Japan. He was a Research Scientist at Toshiba Corporation from 2012 to 2020, and an Algorithm Director at Alibaba Group from 2020 to 2023. He has been a Professor at Xi'an Jiaotong University, Xi'an, China, since 2023. He has published one English monograph, one textbook, and more than 20 papers in journals and conferences. His research interests include large language models, embodied AI, speech processing, and natural language understanding.

Nan Li received the Master's degree from Beijing University of Posts and Telecommunications, Beijing, China. He currently serves as Director of the Wireless and Terminal Technology Research Department at China Mobile Research Institute. He has been engaged in research, standardization, and commercial deployment of 5G/5G-Advanced (5G-A) and 6G key technologies, as well as the development of communication chips and integrated communication-computing-intelligence base stations. He has published over 30 papers in scientific journals and conferences and co-authored one technical book. His contributions bridge cutting-edge theoretical advancements with practical innovations in next-generation mobile networks.

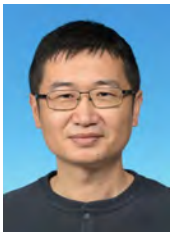
Yong Li is currently a Full Professor of the Department of Electronic Engineering, Tsinghua University, Beijing, China. He received the Ph.D. degree in electronic engineering from Tsinghua University in 2012. His research interests include machine learning and data mining, particularly, automatic machine learning and spatial-temporal data mining for urban computing, recommender systems, and knowledge graphs. Dr. Li has served as General Chair, TPC Chair, SPC/TPC Member for several international workshops and conferences, and he is on

the editorial board of two IEEE journals. He has published over 100 papers on first-tier international conferences and journals, including KDD, ICLR, NeurIPS, WWW, UbiComp, SIGIR, AAAI, TKDE, TMC etc, and his papers have total citations more than 34 000.

Tailin Wu is an Assistant Professor of artificial intelligence at the School of Engineering at Westlake University, Hangzhou, China, and PI of AI for Scientific Simulation and Discovery Lab. He earned the Bachelor's degree from Peking University, the Ph.D. in Physics from Massachusetts Institute of Technology, Massachusetts, USA, and Postdoc in Computer Science, Stanford University, Stanford, USA. His research focuses on developing AI methods for scientific simulation, control, and discovery. He has published over 30 papers in top machine learning conferences and top physics journals, and his work has been applied in large-scale simulations of fluids, energy, mechanics, and other fields in physics. He also serves as an Area Chair for top machine learning conferences such as NeurIPS and ICLR, and reviewer for Nature Machine Intelligence, Nature Computational Science, Nature Communications, and Science Advances.

Wei Xu received the B.Sc. degree in electrical engineering and his M.S. and Ph.D. degrees in communication and information engineering from Southeast University, Nanjing, China in 2003, 2006, and 2009, respectively. Between 2009 and 2010, he was a Post-Doctoral Research Fellow at the University of Victoria, Victoria, Canada. He was an Adjunct Professor of the University of Victoria in Canada from 2017 to 2020, and a Distinguished Visiting Fellow of the Royal Academy of Engineering, UK in 2019. He is currently a Professor at Southeast University, Nanjing, China. His research interests include information theory, signal processing, and artificial intelligence for wireless communications.

He received the Science and Technology Award for Young Scholars of the China Institute of Communications in 2018 and the National Natural Science Foundation of China for Outstanding Young Scholars in 2020. His work was recognized with the IEEE Communications Society Heinrich Hertz Award in 2023 and multiple Best Paper Awards at IEEE ICC 2024, IEEE Globecom 2014, etc. He has been named as a Highly Cited Researcher by Clarivate. He served as an Editor for IEEE Transactions on Communications from 2018 to 2023, and an Editor and Senior Editor for IEEE Communications Letters from 2015 to 2023. He is now serving as an Area Editor for IEEE Communications Letters and an Associate Editor for IEEE Transactions on Mobile Computing and IEEE Transactions on Vehicular Technology. He is a Fellow of IEEE and a Fellow of IET.



Jun Zhang received the Ph.D. degree in Electrical and Computer Engineering from the University of Texas at Austin, Austin, USA, in 2009. He is a Professor in the Department of Electronic and Computer Engineering (ECE) and Associate Director of the Computer Engineering (CPEG) Program at The Hong Kong University of Science and Technology, Hong Kong, China. His research interests include integrated communications and AI, generative AI, and edge AI systems. He is an IEEE Fellow and an IEEE ComSoc Distinguished Lecturer (2023-2024). He is a co-recipient of several best paper awards, including the 2025 IEEE Communications Society Katherine Johnson Young Author Best Paper Award, the 2021 Best Survey Paper Award of the IEEE Communications Society, the 2019 IEEE Communications Society & Information Theory Society Joint Paper Award, and the 2016 Marconi Prize Paper Award in Wireless Communications. He also received the 2016 IEEE ComSoc Asia-Pacific Best Young Researcher Award. He is currently an Area Editor of IEEE Transactions on Wireless Communications (leading the area of machine learning and artificial intelligence) and IEEE Transactions on Machine Learning in Communications and Networking (leading the area of distributed learning and AI at the network edge).

Qi Sun received the Ph.D. degree in Information and Communication Engineering from Beijing University of Posts and Telecommunications, Beijing, China. She is a Senior Researcher at China Mobile Research Institute. She has been working on key technologies and standardization of 5G/6G radio access networks, and her current research focuses on the deep convergence of communication, AI, and computing in 5G and 6G networks. She has filed more than 40 patents and published over 40 papers.