

Networking Systems of AI: On the Convergence of Computing and Communications

Liang Song^{ID}, Senior Member, IEEE, Xing Hu^{ID}, Guanhua Zhang, Petros Spachos^{ID}, Senior Member, IEEE, Konstantinos N. Plataniotis^{ID}, Fellow, IEEE, and Hequan Wu

Abstract—Artificial intelligence (AI) and 5G system have been two hot technical areas that are changing the world. On the deep convergence of computing and communication, networking systems of AI (NSAI) is presenting a paradigm shift, where distributed AI becomes immersive in all elements of the network, i.e., cloud, edge, and terminal devices, which make AI virtually operating as a networking system. On the other hand, by the evolution of the communication systems, a network is becoming a service-specific system interwoven with AI, i.e., the network operates as an AI system, enabling real-time smart services. With the developing technology trends of “AI as a network and network as an AI,” the ecosystem of NSAI can be presenting the next-generation waves of both AI systems and B5G-6G communication networks. In this article, we mainly aim to provide a comprehensive survey on the system architecture, key technologies, application scenarios, challenges, and opportunities of NSAI, which can shed light on the future developments of both telecommunications and AI computing. The contributions of this article also include: 1) providing a unified framework for the deep convergence of computing and communications, where the network and application/service can be jointly optimized as a single integrated system and 2) suggesting the roadmap and open research problems in realizing the online-evolutive integration of cyberspace, physical world, and human society, toward the ubiquitous brain networks (UBNs), which are requiring the joint efforts from both research communities of computing and communication.

Index Terms—6G, AI-powered brain-computer interface (ABCI), artificial intelligence (AI), B5G, Internet of Things (IoT), networking systems of AI (NSAI), ubiquitous brain network (UBN).

I. INTRODUCTION

OVER the past decades, the industry has witnessed the fast development of both computing and communication

Manuscript received 21 December 2021; revised 18 March 2022; accepted 28 April 2022. Date of publication 3 May 2022; date of current version 7 October 2022. This work was supported in part by the China Mobile Research Fund of Chinese Ministry of Education under Grant KEH2310029, and in part by the Double First-Class Funds, Fudan University, as well as Shanghai Key Research Lab of NSAI, China. (Corresponding author: Liang Song.)

Liang Song and Guanhua Zhang are with the Academy for Engineering and Technology, Fudan University, Shanghai 200433, China (e-mail: songl@fudan.edu.cn).

Xing Hu is with the School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China.

Petros Spachos is with the School of Engineering, University of Guelph, Guelph, ON N1G 2W1, Canada.

Konstantinos N. Plataniotis is with the Department of ECE, University of Toronto, Toronto, ON M5S 1A1, Canada.

Hequan Wu is with the Chinese Academy of Engineering, Beijing 100088, China.

Digital Object Identifier 10.1109/JIOT.2022.3172270

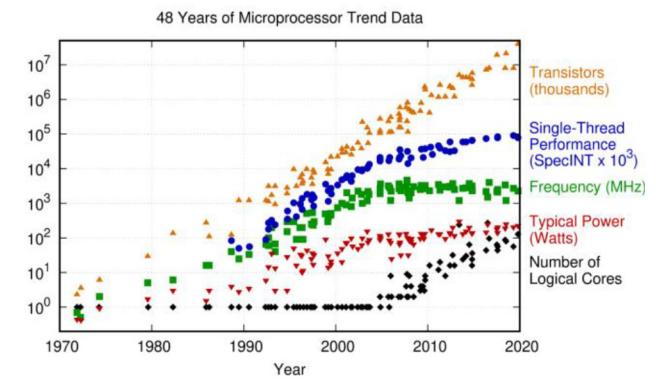


Fig. 1. Growth of computing over time [1].

technologies. On the computing side, as shown in Fig. 1, brilliant engineers have been endeavoring to keep the engineering practice of Moore’s law, which exponentially improves the capabilities of state-of-the-art computers over time. Today’s mobile devices, e.g., smartphones, have much higher computing power than desktop computers of decades ago, and consume much less power at the same time. A variety of lightweight devices/terminals with diverse computing and communication capabilities are getting popular with the development of the Internet of Things (IoT) beyond smartphones, such as unmanned aerial vehicles (UAVs), iPads, smart vehicles, wearable computing, including smartwatches and glasses, servicing robots, etc. On the communication side, as shown in Fig. 2, the global industry has been migrating to a new generation in almost every decade, where every new generation has contributed to a revolution of the whole telecommunication ecosystem. Today, people expect round-the-clock connectivity at home, work, and everywhere, with seamless access to services, such as video streaming, Internet shopping, and video conferencing wherever and whenever needed. The result is a simpler, faster, and more agile network that can enable a seamless transition from wired to wireless, and *vice versa*, delivering uninterrupted, omnipresent connectivity to customers at lower cost. During the ongoing wired and wireless communication convergence and network virtualization procedure, mobile communications have been playing a key role in improving the end-to-end Quality of Service (QoS). Besides the 10–100 times improvement in communication metrics over every generation, e.g., in throughput, capacity, latency, power efficiency, etc., the migration from 2G to 3G/4G virtually transformed traditional computer

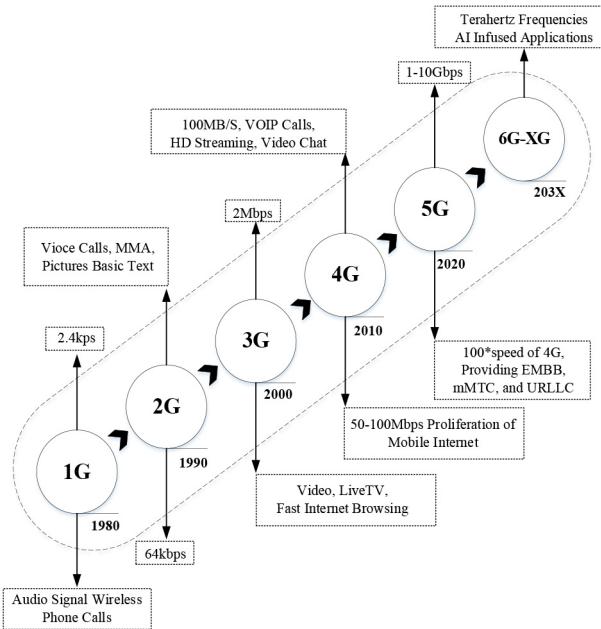


Fig. 2. Migration of global telecommunications.

networks into today's mobile Internet. The latest evolution to 5G provides three primary sets of application scenarios, i.e., enhanced mobile broadband (eMBB), massive machine-type communications (mMTCs), ultrareliable low latency communications (URLLCs), which are leading to the next generation of the Internet of Everything (IoE), connecting every smart mobile/edge device in numerous verticals, such as urban transportation, medical, industrial, video production, etc.

Presently, artificial intelligence (AI) has been in its third-wave prosperity since the popularity of deep learning methodologies [2]. AI has achieved great success in areas including computing vision [3], natural language processing [4], multimedia production [5], medical and medicine applications [6], and even human genetics [7], among others. In applying AI to communication networks, for example, Li *et al.* [8] presented the opportunities and challenges to exploit AI for achieving intelligent 5G networks and demonstrated the effectiveness of AI to manage cellular network resources. Shafin *et al.* [9] argued that deploying AI in 5G and beyond would require surmounting significant technical barriers in terms of robustness, performance, and complexity, and presented a possible roadmap to realize the vision of AI-enabled cellular networks for the Beyond-5G (B5G) and the sixth-generation (6G) infrastructure.

The convergence of computing and communications presents exciting opportunities to both technical communities. From the computing perspective, state-of-the-art AI engineering requires many manual steps and suffers from bottleneck issues such as "generalization." AI algorithms have been designed and trained in a centralized manner, which require engineers to build a labeled data set and a predetermined algorithmic network, e.g., neural networks. Problems arise when the training set is insufficient, or attributes of the training set do not match well with real-world data. Moreover, as traditional AI is maintained in a data center, e.g., cloud computing,

its response time can be insufficient to real-world situations, e.g., in autonomous driving, which requires a strong case of interaction. By embracing the new generation of communication networks, it is understood that AI will be populated in the entire networking system of cloud-edge-device [10]. It not only provides higher user data privacy [11] but also breaks the current AI bottlenecks by making it much "faster" and "smarter." Particularly, by deploying AI on the edge and terminal devices, it can better interact with humans and the physical world in real time; and by adopting local data and multiagent unsupervised learning [12], the next-generation AI will be more adapting to the changing environment and users without manual intervention, which can effectively lay down the principal foundations of "online-evolutive learning" and generic AI [13], [14].

From communications and network engineering perspectives, beyond taking AI as a tool to better manage communication-network resources, the telecommunication community is also embracing AI as the potential killer applications for 5G and potentially 6G networks. Distributed AI computing in many application scenarios is requiring numerous terminal devices to exchange a huge amount of data or model parameters with strict latency requirements, known as the data and model synchronization [15]. Considering the dynamics of applications and computing, distributed AI is a perfect example that requires smart services with QoS assurance.

On the deep convergence of computing and communication, we aim to provide a new paradigm shift on the networking systems of AI (NSAI), by overviewing state-of-the-art works. In NSAI, distributed AI becomes immersive in all elements of the network, i.e., cloud, edge, terminal devices, which make AI virtually operating as a networking system. On the other hand, by the evolution of software-defined networks/network function virtualization (SDN/NFV) [16], a network is becoming a service-specific system interwoven with AI, i.e., the network operates as an AI system, enabling the real-time smart services. With the developing technology trends of *AI as a network system* and *network as an AI system, vice versa*, the ecosystem of NSAI can be presenting the next-generation waves of both AI systems and telecommunication networks.

Fig. 3 illustrates the structure of this article. The main contributions of this work include the following.

- 1) Providing a comprehensive survey on the system architecture, key technologies, application scenarios, challenges, and opportunities of NSAI.
- 2) Proposing a unified framework for the deep convergence of computing and communications, where the network and application/service can be jointly optimized as one integrated system.
- 3) Suggesting the roadmap and open research problems in realizing the online-evolutive integration of cyberspace, physical world, and human society, toward the ubiquitous brain networks (UBNs).

This article is organized as what follows: the visions and technical architecture of NSAI are reviewed in Section II; the key technologies are reviewed in Section III; the application scenarios and services of NSAI are described in Section IV,

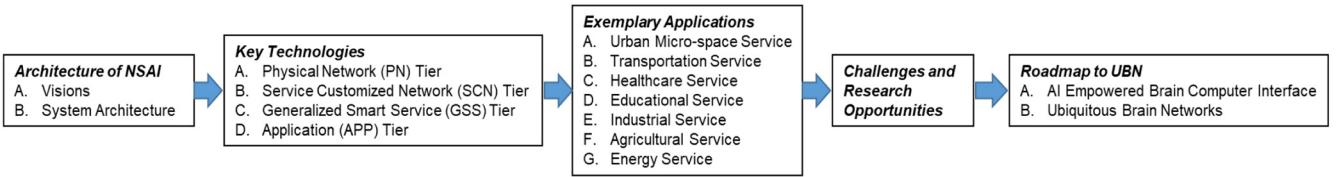


Fig. 3. Paper structure.

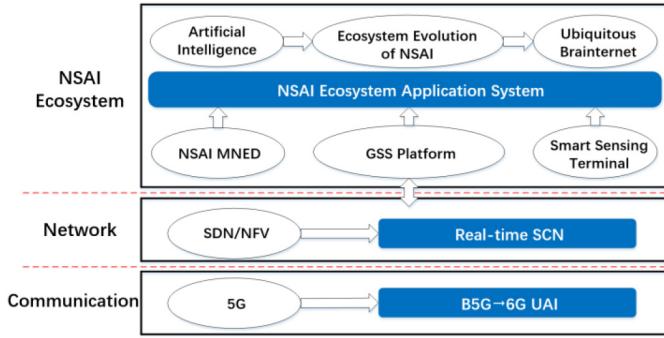


Fig. 4. Visions of NSAI, including SCN, GSS, UAI, MNED, and the evolution of the ecosystem.

which expects to merge cyber–physical spaces and human society; existing technical challenges and opportunities are then summarized in Section V; the developing roadmap of NSAI is provided in Section VI to have human intelligence an integral part of NSAI, i.e., a roadmap to UBN—a futuristic merger of artificial and human intelligence; finally, concluding remarks are given in Section VII.

II. ARCHITECTURE OF NSAI

A. Visions

With the convergence of computing and communications, NSAI shall at least be service customized, where the network resources can be customized for specific smart service provision. Moreover, since traditional Internet Protocols (IPs), e.g., TCP/IP, are insufficient in handling network dynamics, such as mobility, topology changes, and resource reallocation [17], by the evolution of SDN/NFV, a dynamic service-customized network can be created virtually for every smart service, and new networking protocols shall be designed upon the virtual service network to support distributed AI computation. Visions of NSAI are illustrated in Fig. 4, which are further described in what follows.

- 1) *Service-Customized Network (SCN)*: A service-customized virtual network expects to provide enough configurability for dynamic service adaption. For example, in smart transportation services, the number of vehicles/terminals in rush hours can be much higher than rest time of the day, which constitutes different scales of network QoS requirements and network resources. The virtual network topology and the network resources of computing/communication/caching shall be dynamically reconfigured to accommodate the actual changes of service requirements. With further network

slicing to accommodate different users of the service, SCN can be an evolution of SDN/NFV, providing a framework for NSAI to real-timely create, configure, reconfigure, and slice an AI-embedded network by multiple time scales.

- 2) *Generalized Smart Service (GSS)*: With the integration of typical applications and services, GSS is expected to provide a set of generalized application programming interface (API) for interaction with NSAI applications. Particularly, the complexity of communications and computing shall be hidden from application developers. The entire virtual network becomes a massive computing entity with online-evolutive learning, including everything connected, such as all the smart terminal devices, edge, and cloud computing servers in the SCN. The heterogeneous and massive data are going to be processed real timely in GSS, and then to be supplied to the NSAI applications. GSS can dynamically reconfigure both the SCN and applications in providing reliable and real-time services when both network conditions and application requirements can be soft and changing over time.
- 3) *B5G Unified Air Interface (UAI)*: Although 5G has defined eMBB, mMTC, and URLLC as three sets of vertical applications, the 5G system (5GS) was not designed to support high throughput, massive connections, and low latency at the same time. With the future prolific NSAI applications and GSS, it is expected that requirements of eMBB, mMTC, and URLLC will merge in SCN, especially for generic application scenarios, such as smart and healthy cities (SHCity), or smart infrastructures in general, where every smart device can be interconnected in the service area for sophisticated applications. From B5G to 6G, a new air interface is expected to unify eMBB, mMTC, and URLLC, being dynamically reconfigurable [18] under SCN.
- 4) *Micro and Nano Electronic Devices (MNEDs)*: It is expected that state-of-the-art micro and nanoelectronics technology will still be fast developing, so that more AI and communication capabilities can be fitted in NSAI terminal devices, with even lower power consumption. This is enabled by the continuity of Moore’s law in nanoscale, and the new 3-D chip architecture that can potentially integrate computing, caching, sensing, and communications. As compared to the human brain, it is estimated that state-of-the-art electronics are still over one million times less power efficient in supporting deep neural networks, let alone the scale/size difference, etc. This also indicates a huge potential room for

TABLE I
EXAMPLES OF INDUSTRIAL WORKS RELATED TO NSAI

Qualcomm	Snapdragon 768G mobile platform [19], is designed to bring a next-level performance that enables smart, immersive gaming experiences with the integration of truly global 5G, sophisticated on-device AI, and selected Qualcomm Snapdragon Elite Gaming features.
ZTE	uSmartInsight 2.0 platform [20], is a converged platform integrating big data and artificial intelligence. The 2.0 platform is equipped with a distributed training engine developed by ZTE to support higher parallel training speedup. It supports automated distribution while satisfying the major scenarios of model training and model inference. Meanwhile, it enables simple and efficient visual modeling so that end-to-end AI applications can be easily developed and deployed.
HUAWEI	Ascend [21] is Huawei's full-stack all-scenario AI solution. Full-stack means the outstanding, cost-effective computing power and low-barrier application development platform to AI application developers. This platform implements AI data modeling and model training, making application development simpler, more agile, and more efficient. All scenarios suggest ubiquitous intelligence in all commercial scenarios across device, edge, and cloud.
Nokia	AVA 5G cognitive operations platform [22], is built to monitor and manage the networking requirements of 5G networks. Traditional management methods are no longer suitable for 5G networks since larger volumes of traffic and new edge technologies are required. The monitoring platform automates network functions and can detect faults up to a week in advance, giving carriers ample time to fix the issues and avoid costly network failures.
Ericsson	Ericsson Network Intelligence and Omni Network Channel [23], enables communication service providers to secure always-on networks and deliver optimal user experiences with the employment of AI, automation, and predictive analytics to address the complex reality faced by communications service providers, including the exponential data growth and the continuous introduction of new technologies such as 5G, digital transformation, and scattered information sources and insights.
Orange	Orange [24] has long been talking up the potential ways in which artificial intelligence (AI) can be put to good use, but now the French giant, which identified AI-enabled innovation as one of the four ambitions of its Engage 2025 strategy, has provided an update on exactly how it can better manage and monitor its networks and services to good effect.
Intel	Intel Neural Network Processors (NNP), the processor [25] for training (NNP-T1000) and inference (NNP-I1000) — is Intel's first purpose-built ASICs for complex deep learning with incredible scale and efficiency for cloud and data center customers. Intel also revealed its next-generation Intel Movidius Myriad Vision Processing Unit (VPU) for edge media, computer vision and inference applications.
China Telecom	The construction of China Telecom's intent-based network [26] follows the overall layout of China Telecom's AI development. The network consists of four parts: intelligent brain, orchestration and control layer, and intelligent infrastructure of the intent-based network, and AI terminals.
AT&T	Google Cloud and AT&T announced [27] a collaboration to help enterprises take advantage of Google Cloud's technologies and capabilities using AT&T network connectivity at the edge, including 5G. These edge computing solutions will be powered by AT&T's network and will utilize Google Cloud's core capabilities in Kubernetes, AI and machine learning, data and analytics, and other leading technologies delivered across a global footprint. By bringing Google Cloud compute and capabilities to the edge, businesses can move infrastructure from centralized locations to these edges and run applications closer to end-users, thereby minimizing latency, optimizing operations, providing stronger security and delivering compelling, innovative end-user experiences.
Vodafone	Vodafone's AI Framework [28] was published in April 2019, with the focus on six areas: transparency and accountability; ethics and fairness; the preservation of privacy and security; human rights, diversity and inclusivity. It maximizes the benefits of AI while managing the disruption of its implementation.
T-mobile	From AI to VR and Beyond: T-Mobile Accelerator Named Class of 2020 Startups [29], is driving the development in AI, drones, robotics, autonomous vehicles and more on the carrier's nationwide 5G/4G networks.
China Mobile	China Mobile, Haier and Huawei jointly [30] launched the world's first AI+5G interconnected factory at the 2019 World Industrial Internet Conference (WIIC) in Qingdao. This improvement refines future smart manufacturing by innovating and transforming enterprise organizations, business models, and ICT technologies, and integrating key technologies of AI and 5G.

improvement in electronics, to realize low-power and online AI evolution on ubiquitous smart mobile devices.

- 5) *Evolution of NSAI Ecosystem:* With the convergence of AI and network to NSAI, it is expected that every AI-empowered device will be connected and transformed to an integral element of a massive and ubiquitous NSAI entity. NSAI will be entering every aspect of the human society, with the online-evolutive learning to provide better and customized user experiences, e.g., in communities, urban areas, transportation, factory, medical and educational services, suburban and agriculture, energy service domain, etc. As 3G/4G created mobile Internet and had cyberspace interacting with the physical world, also known in the industry as “online-to-offline,” NSAI with the 5G/6G will seamlessly integrate the cyber and physical world with human society, which indeed leads to the next generation of the Internet. With the latest development of AI-empowered brain-computer interface (BCI), NSAI shall be contributing to the merger of artificial and human intelligence, which is introduced in this article as UBN, or “Brainternet.”

Visions of NSAI differentiate from state of the art, where the joint optimization of the network and application/service in one integrated system is required in NSAI. On the contrary, traditional studies typically treat application and network as two separated systems. Therefore, NSAI can conceptually present a paradigm shift of “AI as a network system, and network as an AI system,” instead of the traditional “network for AI, or AI for network.” Table I further shows state-of-the-art works of numerous companies in the industry, as related to the aforementioned visions.

B. System Architecture

Numerous pioneering researchers have been trying to define the architecture of AI-empowered networks, which can be generally categorized as AI for network, and network for AI, respectively. For example, Yao *et al.* [31] proposed an intelligent architecture, called NetworkAI, for self-learning control strategies in SDN, which consists of three planes called the forwarding plane, the control plane, and the AI plane. Dai *et al.* [32] proposed an architecture that can dynamically

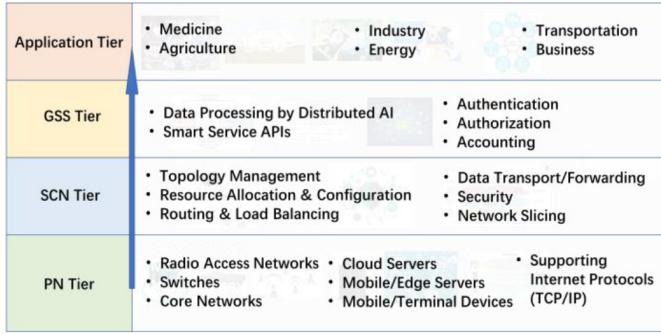


Fig. 5. Architecture of NSAI, including PN, SCN, GSS, and APP tiers.

orchestrate edge computing and caching resources to improve system utility by making full use of AI-based algorithms. Zhao *et al.* [33] proposed an software-defined optical networks (SDONs) which can support the control-layer AI and onboard AI simultaneously, where onboard AI was proposed to be based on edge computing to support various machine learning applications. Guo *et al.* [34] proposed an AI-based semantic IoT (AI-SIoT) hybrid service architecture to integrate heterogeneous IoT devices in supporting intelligent services, where the proposed architecture is empowered by semantic and AI technologies, enabling flexible connections among heterogeneous devices. Rafique and Velasco [35] reviewed several machine learning concepts tailored to the optical networking industry and discussed algorithm choices, data and model management strategies, as well as the integration into existing network control and management tools; four networking case studies were provided in detail, covering predictive maintenance, virtual network topology management, capacity optimization, and optical spectral analysis. Stergiou *et al.* [36] proposed an innovative infrastructure of secure scenario which operates in a wireless 6G network for managing big data on smart buildings. Wang *et al.* [37] proposed a distributed cognitive cellular network, which integrated AI and cognitive radio (CR) technology into a sophisticated multiagent system (MAS). The hierarchical distributed networking framework based on 5G cellular networks was proposed with four tiers: 1) CR users; 2) cellular BSs; 3) cloud processing; and 4) applications.

Given the visions of NSAI and the pioneering research on integrating AI with the network, a generalized NSAI framework not only needs to accommodate state-of-the-art research in both computing and communication but also shall support the joint optimization of network and application/system in a single integrated system. We hereby propose that the system architecture of NSAI can be divided into four tiers, including: 1) physical network (PN) tier; 2) SCN tier; 3) GSS tier; and 4) application (APP) tier. The proposed architecture is shown in Fig. 5.

1) *The PN Tier* consists of the heterogeneous optical and wireless infrastructures, including, for example, radio access networks, switches, core networks, cloud servers, mobile/edge servers, and mobile/terminal devices. The PN Tier can also consist of infrastructures from different operators that work together to provide unified smart services; current IPs, e.g., TCP/IP, can be supported in

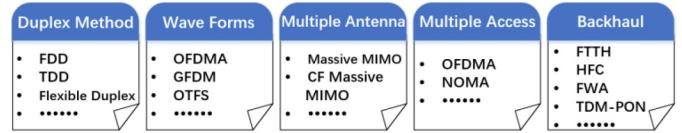


Fig. 6. Recent advances of wireless and wireline interfaces for PN.

the PN Tier with backward compatibility, as an underlay network infrastructure for the upper SCN Tier.

- 2) *The SCN Tier* creates and configures the service-customized virtual network, which shall include functional abstractions, such as topology management, resource allocation, data transport and forwarding, routing and load balancing, security, and network slicing. As an overlay network customized for smart service, SCN can be constructed upon a virtual link layer (L2) and provide backward compatibility to upper tiers.
- 3) *The GSS Tier* mainly handles in-network data processing by distributed AI on the data plane, while it also dynamically configures the SCN and applications on the control plane. Therefore, the network and application can be jointly optimized to enable real-time smart services, where both network resources and application requirements can be soft and changing over time. As presented in the NSAI visions, the GSS Tier works as a middleware platform that seals the complexity of network and data computing from the NSAI applications. The GSS Tier also offers authentication, authorization, and accounting (AAA) for subscribed users of smart services.
- 4) On top of the GSS Tier and its APIs, the NSAI APP Tier is programmed to provide ground-breaking user experiences to consumers, businesses, and municipalities, by creating new Internet horizons beyond the current 4G/mobile Internet and merging the cyberspace with the physical world and human society.

III. KEY TECHNOLOGIES

With the vision and technical architecture of NSAI, we hereby aim to provide a comprehensive review of state-of-the-art technologies in the four individual tiers of NSAI. The development of these technologies can also contribute to a roadmap for the evolution of NSAI.

A. Physical Network Tier

The PN Tier converges the latest developments of wireless and wireline PNs, including the physical interface technologies, new multihop transmission technologies, as well as micro and nanoelectronics. As classical Shannon capacity [38] has only specified the limit of point-to-point linkage, the next-generation PN shall aim at breaking the classical Shannon limit, by exploring the achievable end-to-end capacity of massive and heterogeneous PNs.

1) *Wireless and Wireline Interfaces:* As illustrated in Fig. 6, for the next-generation wireless interface, many kinds of orthogonal frequency-division multiplexing (OFDM) waveform schemes have been proposed in 5G standards [39],

including the multicarrier system with sub-band filtering and subcarrier filtering [40]. The band above 52.6 GHz is considered as a potential band of 6G systems. Frequencies above 52.6 GHz can face more difficult challenges, such as higher phase noise, extreme propagation loss due to high atmospheric absorption, lower power amplifier efficiency, and stronger power-spectrum density regulatory requirements. Orthogonal time–frequency space (OTFS) modulation was proposed in [41] and [42], which modulates the signal to the delay Doppler domain, and equivalently transforms the time-varying multipath channel to the time-delay-Doppler domain. New waveforms, e.g., (windowed) OFDM, GFDM, or OTFS, are being proposed to use in combination with high modulation cardinality, e.g., 256 QAM, and with multiple antennas, to maximize the spectrum efficiency. In multiple access techniques, nonorthogonal multiple access (NOMA) still presents new directions to improve spectrum efficiency toward 6G. Ding *et al.* [43] provided a systematic review of emerging NOMA technologies, from its combination with MIMO to cooperative NOMA, as well as the interplay between NOMA and CR. Gui *et al.* [44] proposed an effective deep-learning-aided NOMA system, in which several NOMA users with random deployment are served by one base station. Zhang *et al.* [45] proposed a scheme where evolved NodeB (eNB) can distinguish the multiplexing user equipments (UEs) in the power domain. With the potential high density of next-generation radios, cell-free (CF) massive MIMO was proposed to overcome the intercell interferences [46], where the performance gain comes from the joint processing of a large number of geographically distributed remote antenna units (RAUs) [47]. CF massive MIMO has been shown to achieve higher spectrum efficiency than centralized massive MIMO [48], and [49] provided a comprehensive survey of CF massive MIMO systems. Buzzi and D’Andrea [50] introduced a user-centric virtual cell approach to CF massive MIMO, wherein each user was served only by a limited number of access points. Zhang *et al.* [51] proposed a framework for the performance analysis in the CF massive MIMO with classical hardware distortion models, where closed-form equations for spectrum and energy efficiency were derived.

To accommodate massive connections and mobile traffics, the optical transport network has been playing an important role in carrying 5G radio signals. The optical module is an important device in the 5G optical transport network. Since wireline transports are used with different rate requirements in fronthaul, midhaul, and backhaul, different types of optical modules are required with various challenges. Ji *et al.* [52] surveyed the challenges, recent studies, and potential solutions for the 5G flexible optical transport networks with the performance requirements on large capacity, low latency, and high efficiency. Currently, dynamic bandwidth allocation mechanism is becoming an important part of time-division-multiplexing passive optical network (TDM-PON) for fronthaul; and ultralow latency transport is required for designing a novel optical system under more stringent latency constraints. For example, Zhou *et al.* [53] demonstrated a novel mobile fronthaul architecture based on functional split and TDM-PONs with a unified mobile and PON scheduler

known as Mobile-PON. Waqar *et al.* [54] evaluated the performance of PON-based links for long-distance communications in fronthaul systems. Mikaeil *et al.* [55] proposed an optimized dynamic bandwidth allocation known as optimized round-robin, supporting fronthauling over 10G-passive optical networks (XG-PONs). Nishitani *et al.* [56] studied Type B protection for a TDM-PON system, and Cano *et al.* [57] performed an experimental demonstration of a statistical OFDM-PON with multiband optical network units (ONUs) and elastic bandwidth allocation. Additionally, compared to wireline access technologies, such as fiber-to-the-home (FTTH) or hybrid fiber-coaxial (HFC), fixed wireless access (FWA) has also presented several benefits [58]. FWA’s rollout cost can be significantly lower than the wireline access technologies, by eliminating the last mile fiber or cable reaching the customer premises. FWA can also be much quicker than other wireline access technologies, whose deployments usually involve complex processes to obtain permits for installing wires.

2) *New Multihop Transmission:* Traditional wireless communications have been limited to the last hop in cellular infrastructures. However, in the recent development of 5G standards [39], L2 multihop technologies have been introduced, in NR-IAB and NR-sidelink, which virtually covers all linkages including uplink, downlink, and sidelink in cellular networks. Cognitive L2 multihop communications were first proposed in [59]–[61], where radio resources, including both spectrum and radio stations are opportunistically utilized in the multihop linkage to realize *reliable end-to-end throughput and latency over any number of wireless hops*. It was further studied in research works, such as [62]–[67] from numerous perspectives, including node deployments, mobility, and applications. From other perspectives, Tin *et al.* [68] considered a cooperative multihop secured transmission protocol, where a secondary source attempts to transmit its data to a secondary destination with the assistance of multiple secondary relays. Goel *et al.* [69] analyzed the performance of relay selection at each hop in a decode-and-forward-based multihop CR system with cooperative spectrum sensing. Jiang *et al.* [70] studied the collaborative multihop routing in cognitive networks and constructed the collaborative routing in multihop cognitive networks by considering the interference among nodes. Sheikholeslami *et al.* [71] considered multihop covert communication over a moderate size network where the relays can transmit covertly by using either a single key for all relays or different independent keys at the relays. El-Banna *et al.* [72] proposed a machine-learning-based selection approach that adaptively chooses the best forwarding scheme in hybrid multihop dense networks. Majid *et al.* [73] developed a discrete-component-based backscatter tag-to-tag transceiver and a communication protocol suite. Ying and Nayak [74] exploited the social relationships from interactions and contributions, and then formulated an optimal relay selection problem to enhance cooperative multihop device-to-device (D2D) communications. Challita and Saad [75] proposed a backhaul scheme that relies on UAVs as an on-demand flying network. To address the problem of various link rates between the receiving and sending sides of the relay node, Fadlullah *et al.* [76] considered the channel conditions of

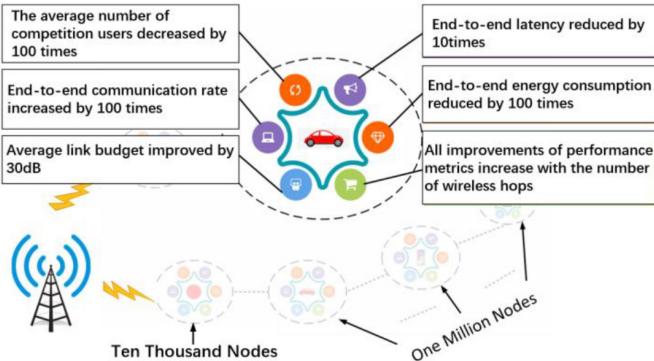


Fig. 7. Advantages of new multihop transmissions (example of ten hops).

each band on either side of the relay, where the receiving and sending rates at the relay node are measured and calculated, respectively.

It was understood from network information theory that the capacity of a massive wireless network can be increasing with the number of nodes (radio stations) N , as compared to point-to-point Shannon capacity, in terms of $O(\sqrt{N})$ [77], and $O(N)$ [78], by considering multihop linkage and node mobility. As illustrated in Fig. 7, reliable multihop communications can largely improve the end-to-end throughput, latency, and energy efficiency, where the performance margin increases at least on the order of the number of wireless hops. It remains to be identified how the new multihop transmissions can contribute to breaking the classical Shannon limits in the next-generation PNs.

3) *Micro and Nano Electronics*: As shown in Fig. 8, state-of-the-art AI electronics are mainly divided into general chips and special chips in terms of technical architecture. The general-purpose chip design refers to the traditional chip architecture, which supports deep learning and complex neural network algorithm through software programming, mainly including CPU, GPU, DSP, FPGA, etc. Although benefiting from mature architectures, general-purpose chips suffer from inefficiency due to, for example, the separation of operation and storage. Even if the complex operation can be realized through software programming, the bandwidth limitation and the consequent power consumption problem during memory access are limiting its performance [79]. Special chip design, also known as application specific integrated circuits (ASICs), provides alternative architectures tailored for AI algorithms. TPU chip [80] introduced by Google can be used for both neural network training and reasoning. At present, TPU has achieved more than 100 PFLOPS processing capacity, and has established the air cycle of chip design, logic implementation, platform mode, and application environment. Cambricon [81], as introduced by Cambrian Science and Technology, uses RISC inspired instruction set to decompose complex neural network computing into modules, to design simple and short call instructions simplify chip design, and reserve space to support future variable algorithms. The true north brain chip [82] proposed by IBM can simulate the human brain computing process, which achieves higher speed computing with very low energy consumption.

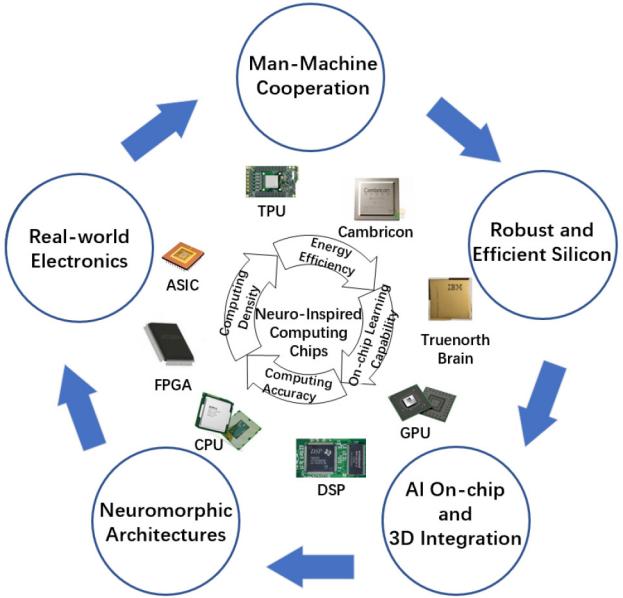


Fig. 8. Micro and nano electronics.

In integrating AI computing, sensing, caching and communication at micro and even nano 3-D scale [83], the current research is still at its infant stage. For example, Liu *et al.* [84] demonstrated an on-chip nano-electronic-based chemical signaling technique for the investigation and improvement of sustainable energy technologies. Sarkar *et al.* [85] presented a straintronics-based nanorobotic system and its VLSI architecture. Straintronics or strain-based spin electronics is an emerging field of IC fabrication technology for modern and future electronics. James [86] designed an integrated co-processor chip based on a memristor crossbar array and complementary metal-oxide-semiconductor (CMOS) control circuitry that can be used to implement neuromorphic and machine learning algorithms. Chen *et al.* [87] designed a fully integrated non-volatile computing-in-memory structure that offers high energy efficiency and low latency for Boolean logic and multiply-and-accumulation (MAC) operations. The approach offered an access time of 4.9 ns for three-input Boolean logic operations, a MAC computing time of 14.8 ns and an energy efficiency of 16.95 tera operations per second per watt. Applied to a deep neural network using a split binary-input ternary-weighted model, the system can achieve an inference accuracy of 98.8% on the MNIST data set. Zhang *et al.* [88] proposed four key metrics for benchmarking neuro-inspired computing chips, i.e., computing density, energy efficiency, computing accuracy, and on-chip learning capability, and discussed co-design principles, from the device to the algorithm level, for neuro-inspired computing chips based on nonvolatile memory. Murmann and Hoefflinger [89] proposed five major thrust areas of nano-chips: robust and efficient silicon, real-world electronics, neuromorphic architectures, AI on-chip and 3-D integration, and man-machine cooperation.

B. Service-Customized Network Tier

The SCN Tier is developed based on the evolution of SDN/NFV and provides a dynamic service-specific virtual

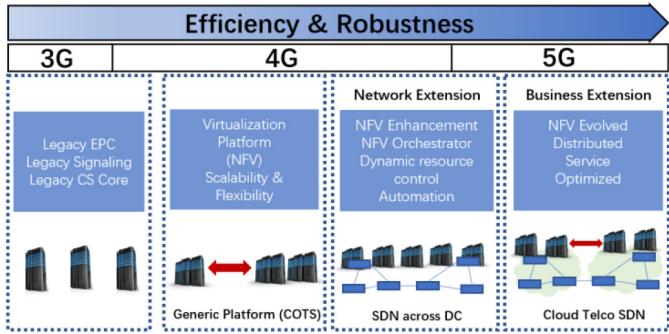


Fig. 9. Evolution of SDN/NFV.

network on top of heterogeneous wireless and wireline PNs. SCN also provides reconfigurable network functionalities, such as network slicing, topology management, resource allocation, load balancing, routing/transporing, security, etc.

1) Evolution of SDN/NFV: As illustrated in Fig. 9, SDN [90], [91] has become a mainstream paradigm that separates the network's control logic from the underlying routers and switches, promoting the logical centralization of network control and the ability to program the network [92]. SDN architecture consists of four planes at a high level, i.e., data plane, controller plane, application plane, management and administration plane [17], [93]–[95]. NFV is closely related to SDN, for example, Li and Chen [96] presented a thorough review on the development of NFV under the software-defined NFV architecture, with an emphasis on service chaining and its application. The NFV architecture is mainly composed of three main components: NFV Infrastructure, virtualized network function (VNF), and NFV Management and Orchestration (M&O) [97], [98]. NFV aims to abstract network forwarding and other networking functions from the hardware on which it runs, and virtualize all PNs beneath a hypervisor, which allows the network to grow without the addition of more devices.

SDN/NFV is still a fast-evolving research area [99], to have *enhanced virtual network efficiency and robustness*. For example, Huang *et al.* [100] devised an SDN inside the mobile edge computing (MEC) architecture to tackle the complicated issues of vehicular ad hoc networks (VANETs) offloading. Chaudhary *et al.* [101] proposed an SDN-based big-data management approach concerning the optimized network resource consumption, such as network bandwidth and data storage units. Zhang *et al.* [102] tackled the VNF placement problem by first proposing a general 5G network framework, which jointly contained both edge–cloud and core–cloud servers. Li *et al.* [103] presented the design and implementation of NFV-RT, a system that dynamically manages resources to provide timing guarantees. Li *et al.* [104] proposed a lightweight NFV framework named DeepNFV, which was based on the Docker container running on the network edge, and integrated state-of-the-art deep learning models with NFV containers to address some complicated problems, such as traffic classification, link analysis, etc. Yang *et al.* [105] proposed an extensible SDN and NFV-enabled network traffic monitoring

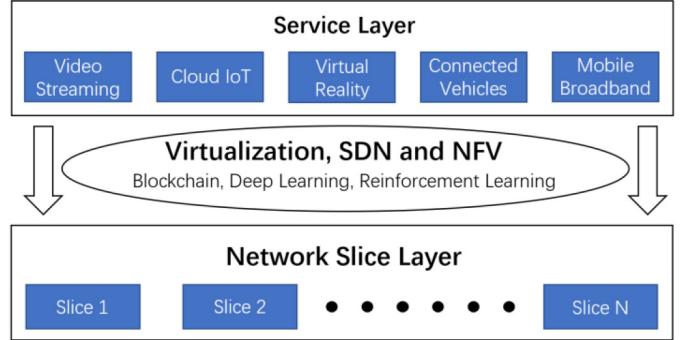


Fig. 10. Network slicing for diversified SLA.

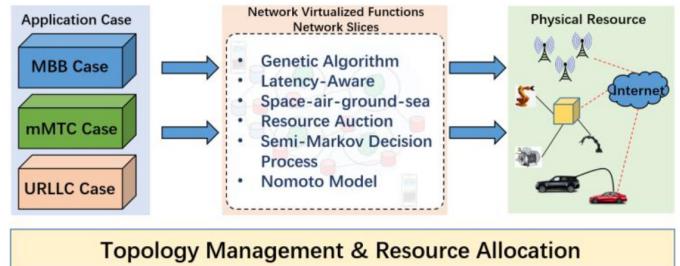


Fig. 11. Virtual network resource allocation.

system, which can closely match the performance of traditional networks at cheaper costs and by adding more flexibility to network management tasks.

2) Improved Network Slicing on Virtual Networks: Network slicing [106] is a sliced network architecture that enables the multiplexing of virtualized and independent logical networks on the same network infrastructure [107]. By using the network slicing technology, one can create a dedicated slice on the common network architecture, where a network slice is a logically isolated end-to-end network tailored for a service type with an agreed service-level agreement (SLA) [108]. Based on the evolution of SDN/NFV, network slicing is also being improved for operating over the service-customized virtual networks, to provide *improved user experiences and diversified SLA* (illustrated in Fig. 10). For example, Choi and Park [109] presented network slice architecture for the 5G core network and provided working mechanisms for slice selection, routing, and slice reselection to support multiple logical networks. Backman *et al.* [110] presented a blockchain slice leasing ledger concept, and the analysis of its applicability in future factories. Bega *et al.* [111] proposed a general framework for AI-based network slice management, introducing AI in the different phases of the slice lifecycle, from admission control to dynamic resource allocation in both core network and radio access network. Sciancalepore *et al.* [112] proposed a reinforcement-learning-based 5G network slice broker which consists of four blocks for supporting network slicing: 1) traffic and user mobility analysis; 2) a learning and forecasting scheme per slice; 3) optimal admission control decisions based on spatial and traffic information; and 4) a reinforcement process to drive the system toward optimal states.

3) *Topology Management and Resource Allocation:* Based on the virtual-network slicing, as shown in Fig. 11, efficient resource allocation schemes have been exploited to improve the flexibility, automation, and intelligence, according to QoS demands of eMBB, URLLC, and mMTC. Moreover, customized smart services can require more precise and efficient resource allocation under dynamic network topology, *e.g.*, for Communications, Computing, and Caching (3C). With the evolving SDN/NFV architecture, network resources can be virtualized and managed in a resource pool [113]. Due to the limited network resources, the increasingly diversified network services and heterogeneous networks, efficient and flexible resource allocation schemes are needed with interference awareness [114]. For example, Zhou *et al.* [115] modeled the resource allocation problem as a semi-Markov decision process, which is defined by state space, action space, reward, and transition probability distribution. The reward function jointly considers the total income, the cost of available resources and the utilization of the total resource for the infrastructure network. Oladejo and Falowo [116] addressed a latency-aware dynamic resource allocation problem for 5G sliced networks in a multitenant multietier heterogeneous environment and proposed a genetic algorithm (GA) intelligent latency-aware resource allocation scheme (GI-LARE). Halabian [117] introduced a distributed solution for the resource allocation problem by forming resource auctions between the slices and the data centers. As future smart services would be provided by a convergent, heterogeneous, virtualized, dynamic and intelligent network, the space–air–ground–sea service [260] has been proposed under specific circumstances, such as maritime, ocean, and desert, and possible to provide with the technology evolved to B5G/6G. Xu *et al.* [118] introduced a space–air–ground–sea integrated network architecture with edge and cloud computing components to provide flexible hybrid computing service for maritime service. In the integrated network, satellites and UAVs provide the users with edge computing services and network access. Yang *et al.* [119] proposed a classical Nomoto model, combined with the actual any slight wind, wave, and current forces on the unmanned surface vehicle, and finally combined with the proportion integration differentiation algorithm to model the surface vehicles.

4) *Other SCN Functionalities and Security:* Based on the dynamic topology and resource management of SCN, the virtual network can embrace novel network functionalities and protocols, which can better support mobility and dynamic network requirements than legacy IPs. The new protocols of SCN can however be integrated with IPs by L2 overlays and underlays [259]. Although functionalities, such as routing, transport, and load balancing were traditionally dealt with separately in multiple layers of network-protocol stacks, they can also be integrally designed, *e.g.*, by the cross-layer design [120]. Research on routing in virtual networks has been trying to make routing more intelligent, in terms of learning abnormal situations, such as congestions, and mobility awareness. For example, Tang *et al.* [121] proposed a new intelligent network flow control method with a self-routing decision based on real-time deep learning, which used a deep convolutional neural network with unique input and output to represent the

backbone of the wireless network. Rusek *et al.* [122] proposed a graph neural network (GNN) model to understand the complex relationship between topology, routing, and input traffics to produce accurate estimates of the per-source/destination pair latency and jitter.

Congestion control has been the focus of transport research. Traditional congestion control methods mainly use end-to-end ACK or NACK feedback to infer the occurrence of network congestion, such as TCP protocols. By analyzing network congestions, the most accurate way to predict congestion is to directly analyze the queue of each node in the routing path, where collaborative modeling and analysis of multiple nodes are needed to achieve early prediction of network congestion. For example, Mao *et al.* [123] proposed to use deep learning to perform large-scale network queuing analysis, where a central node obtains the queue status (including queue size, input communication rate, output communication rate, etc.) at each node, and analyzes the data accumulation of the queue and the congestion status of specific nodes. Pham *et al.* [124] defined a multiple-objective virtual network embedding (VNE) problem called the congestion-aware, energy-aware VNE, which aimed to save cost and energy, and avoid network congestion simultaneously. Stergiou *et al.* [125] studied the use of various opensource tools to achieve a type of network that can provide the more intelligent media-data transfer. Psannis *et al.* [126] proposed an efficient algorithm for advanced scalable media-based smart big data.

To adapt to the explosive growth of data traffic, reasonable and effective load-balancing solutions have also been investigated, which can adapt to dynamic and irregular network topologies, complex interferences, and diverse application requirements. For example, Aghdai *et al.* [127] proposed an in-network congestion-aware load balancer to achieve even load distribution among service instances and optimal network resources usage. Xu *et al.* [128] proposed a mobile load-balancing algorithm based on deep reinforcement learning (DRL) and a two-layer architecture to solve the large-scale load-balancing problem of ultradense networks. Li *et al.* [129] proposed an adaptive and fast redistribution load-balancing strategy and a distributed parallel computing model to solve the chaotic problem caused by the performance difference between nodes. Chakravarthy and Amutha [130] performed load balancing by calculating in advance the capacity of every switch across the packet routing paths.

The concept of resource sharing between virtual networks and network slices can bring many security issues. Aiming at the security difference in 5G network slicing service, Niu *et al.* [131] proposed the network slice trust degree and established a trust degree calculation model, where the network slice manager can truly and effectively calculate network slice trust value. The provision of secure VNF is mandatory to guarantee correct chaining of network functions. Dwiardhika and Tachibana [132] proposed a VNE based on the security level with VNF placement, where some security VNFs were placed to increase the security level of substrate networks. From a secure IoT perspective, Stergiou *et al.* [261] presented a survey of IoT and cloud computing with a focus on the security issues. Stergiou *et al.* [262]

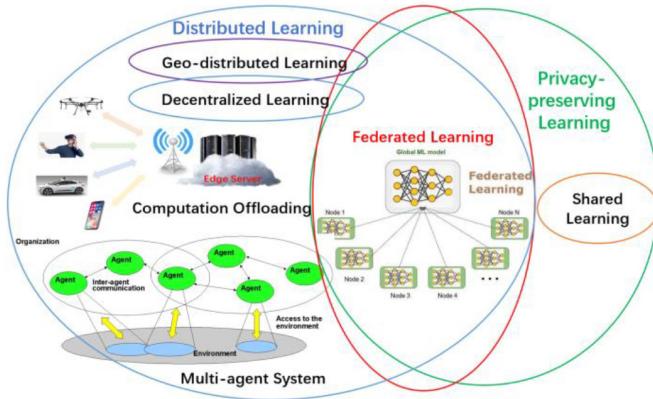


Fig. 12. Distributed AI computing.

proposed a new secure and sustainable system for cloud computing integrated with IoT as a base scenario for big data. The recent development of blockchain technologies also provided a promising framework for *virtual network security and accountability*. Alvarenga *et al.* [133] proposed a blockchain-based architecture for secure management, configuration and migration of VNFs, which ensures: 1) immutability, nonrepudiation, and auditability of the configuration update history; 2) integrity and consistency of stored information; and 3) the anonymity of VNFs, tenants, and configuration information. Rebello *et al.* [134] proposed an NFV-tailored blockchain and a transaction model, called BSec-NFVO, which secured orchestration operations in virtualized networks, ensuring auditability, nonrepudiation, and integrity in a simple and agile way.

C. Generalized Smart Service Tier

The GSS Tier provides distributed AI computing on the data plan and network AAA configurations on the control plane. It also provides a set of service-specific API for the NSAI applications. GSS can operate as middleware of smart services, between the network and applications.

1) *Distributed AI Computing*: In recent years, many types of research have been focused on distributed AI computing for achieving *higher efficiency in learning and inference, user data privacy, and online-evolutive learning*. As illustrated in Fig. 12, distributed AI computing commonly involves how to allocate computing tasks and resources and coordinate every functional module, which needs not only to consider the robustness of data but also to meet the requirements of algorithm precision and generalization [135]–[138].

To protect the privacy of personal data, federated learning has been proposed as a privacy-oriented distributed machine learning approach [139], where raw user data are retained on the user terminal devices. The framework of federated learning consists of the data partition, machine learning model, privacy mechanism, and communication architecture. Data partitioning is mainly operated from two angles of data samples and feature dimensions, where the partitioning based on training samples is mainly divided by random sampling or scrambled segmentation [140]. Machine

learning models can be classified as linear models, decision trees, and neural networks. A multiobjective evolutionary algorithm [141] was designed to minimize the communication cost and global model testing error at the same time, where joint learning is described as a two-objective optimization problem. The communication architecture mainly includes the communication topology based on iterative MapReduce/AllReduce [142], parameter server [143], [144], and data flow. Communication pace is mainly divided into synchronous and asynchronous modes, where algorithms based on synchronous communication mainly include BSP-SGD [145], ADMM [146], EA-SGD [147], etc., and asynchronous communication algorithms mainly include HogWild [148], Cyclades [149], etc. Some researchers also proposed semisynchronous methods, e.g., the stale synchronous parallel (SSP) model [150]. For improving the communication efficiency of federated learning, Konen *et al.* [151] proposed two methods of structural update and sketch update, to reduce the average communication cost. Due to the heterogeneity of computing and communication resources, many kinds of research have been performed to optimize the allocation of federated learning resources. For example, the ELFISH framework [152] analyzed the computing consumption of model training from the aspects of time cost, memory usage, and computing workload. To strike a balance between energy cost and training time, FEDL [153] was designed for every terminal device to approximately solve its local problems up to a certain local accuracy level. The number of devices involved in federated learning can often be very large, and simply minimizing the average loss may not be suitable for the model performance required on some devices. The fair-oriented target Q -fair federated learning (Q-FFL) joint resource optimization [154] was proposed by jointly considering utility function and fairness. Q-FFL minimizes the total weighted loss of Q parameterization, where devices with higher loss can have higher relative weight.

With the rapid development of the IoE, embedded computing devices are becoming ubiquitous, and generating a huge amount of data. Although lightweight AI models can be realized on embedded devices, for computationally intensive tasks, many researchers have investigated the segmentation of inference models, and separately scheduled the workload to devices, edge servers, and cloud. Cooperative computing not only completes the task but also contributes to a balance of device energy consumption, server load, transmission, and execution latency [155], [156]. A related research area is known as computation offloading, which can be divided into local execution, total offloading and partial offloading. Local execution has all computing solely executed on a local terminal device. Total offloading schemes, such as DeepDecision [157], connect powerful edge nodes with weak terminal devices, and the deep-learning inference process can be executed at the terminal or the edge side, depending on the tradeoff between reasoning accuracy, reasoning delay, model size, battery power, and network conditions. Partial offloading refers to offloading some computing tasks to the edge-side node/cloud center, which requires the offloading system to support online fine-grained partitioning of tasks and determine how to assign these

TABLE II
COMPARISON OF DIFFERENT TYPES OF MDRL

Types	Advantages	Disadvantages
Analysis of emergent behaviors [161][162]	Having simple principles, and easy to realize	Unable to solve the non-stationary problem in the multi-agent environment, difficult to converge
Learning communication [163]	Establishing explicit channel to learn optimal strategy	Having too many parameters with complex design architecture
Learning cooperation [164]	No communication modeling, efficient, universal, and easy to realize	Having single application environment (unable to cope with a fully competitive environment)
Agents modeling agents [165]	Predicting the behavior of other agents, having strong robustness, suitable for different scenarios	Having high computational complexity, unsuitable for large systems

divided tasks to terminal devices and edge nodes. For example, the IONN system [158] divides the model into multiple partitions and offloads them to the edge nodes according to certain policies. The edge nodes receive the partition model and build the model incrementally when each partition model arrives.

MAS [159] is also playing an important role in NSAI, e.g., for multi robot cooperation, distributed control, resource management, collaborative decision making, and so on. Typical multiagent learning algorithms include multiagent reinforcement learning (MARL) and multiagent DRL (MDRL), which enable multiagent to find the optimal strategy independently [160]. According to the communication modes between agents, current MDRL can be divided into four categories: 1) “analysis of emergent behaviors” [161], [162] applies the single-agent learning algorithm to the multiagent scene, in order to observe whether there will be new behaviors, such as learning to cooperate, compete or generate new languages; 2) “learning communication” [163] has agents to solve cooperative tasks by learning communication protocols; 3) “learning cooperation” [164] has agents to cooperate under incomplete observation without communication; and 4) “agents modeling agents” [165] has agents to complete tasks by interrupting and modeling other agents’ behaviors. The advantages and disadvantages of various MDRL algorithms are shown in Table II.

MRDL can be presenting a foundation for online evolutive learning (OEL), where the NSAI can be online evolutive in terms of learning from the environment and other AI agents online, as a “society” of AI. The state-of-the-art AI is still dominated by supervised learning, requiring expensive data engineering and offline labeling. The OEL of NSAI can be a paradigm shift, which adapts NSAI to newly generated data and nonstable environmental changes in real time. It may also eventually contribute to the generalized AI system that can adapt to the dynamic changes of open and real-world environments.

2) *AI-Empowered Network Configuration:* With the 5G/B5G background, AI is playing an important role to realize intelligent networking, for dynamic and diversified smart services. Network configuration is no longer just a tool to assign parameters or resources to a virtualized service, but more intelligent algorithms are being developed for AI-empowered network configurations. In the future B5G/6G, the hybrid SDN is evolving into a networking architecture (e.g., in SCN) where both centralized and decentralized paradigms coexist and communicate together to configure, control, change, and manage the network behavior, for optimizing network performance and user experience. Through learning features from the big data generated by the wireless network infrastructure and sensor devices, network configurations can be optimized in real time, resulting in better network performance. Under the paradigm of AI-empowered network configuration, new performance metrics such as *configurability*, i.e., *how swift the network can be configured*, needs to be studied in supporting real-time network adaption.

The AI-aided network configuration has been studied for cellular networks [166], [167]. The Autonomous Networks Project from tmForum [168] aims to define fully automated, zero-wait, zero-touch, and zero-trouble innovative network/ICT services for vertical industries, supporting self-configuration, self-healing, self-optimizing, and self-evolving telecom network infrastructures for telecom internal users, i.e., for planning, service/marketing, operations, and management. Autonomous Networks incorporated a simplified network architecture, autonomous domains and automated intelligent business/network operations for the closed-loop control of the digital business, offering the best possible user experience, full lifecycle operation automation, and maximum resource utilization. Shafin *et al.* [9] and Pan *et al.* [178] showed a typical framework of cellular network optimization based on DRL with the help of a physical model. One of the key tasks is to predict users’ future service requests, location, and mobility, and the network conditions in cellular networks. For example, the processing resources of the central unit can achieve more pooling gain according to the predicted traffic fluctuation. In addition, some access resources can be turned off if no traffic is predicted in the corresponding coverage area. Another preferred example of AI-empowered network configuration can be intelligent vehicular networks. The vehicular network brings unprecedented challenges to traditional wireless communications systems due to its strict and diverse QoS requirements, as well as the inherent dynamics in vehicular environments, such as fast time-varying wireless channels and network topology. AI provides a powerful tool to tackle these problems in network configuration [169]. Among other works, Roy *et al.* [170] introduced BISON, a new bioinspired self-organizing solution for automated and efficient PCID configuration in 5G UDN. Ren *et al.* [171] proposed a VNF Dynamic Auto Scaling Algorithm considering the tradeoff between performance and operation cost, and developed an analytical model to quantify the tradeoff and validate the analysis through extensive simulations. Pérez *et al.* [172] proposed a 5G-oriented solution for proactively detecting and mitigating Botnets, which is a widely used testing network in highly

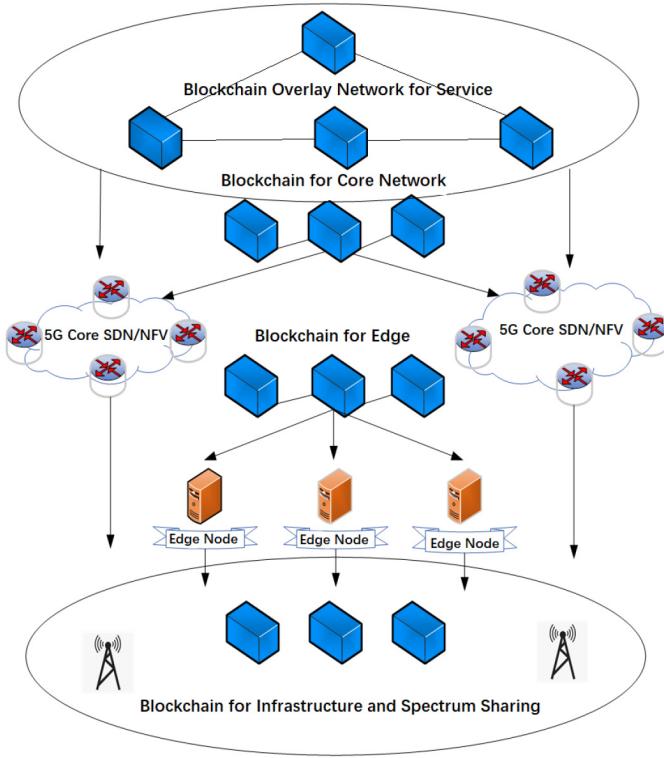


Fig. 13. Blockchain for AAA in NSAI.

dynamic 5G networks. Bega *et al.* [173] presented DeepCog, a deep neural network architecture inspired by advances in image processing and trained via a dedicated loss function. Unlike traditional traffic volume predictors, DeepCog returns a cost-aware capacity forecast, which can be directly used by operators to take short-term and long-term reallocation decisions that maximize their revenues. Maksymyuk *et al.* [174] proposed a deep-learning-based mobile network management and configuration solution for 5G/B5G. Lee *et al.* [175] used a deep learning framework to solve distributed nonconvex constrained optimizations in wireless networks, where multiple computing nodes, interconnected via backhaul links, operated to determine an efficient assignment of their states based on local observations.

3) *Evolution of AAA:* AAA has been a framework for controlling resource access, enforcing policies to use those resources, and auditing their usage. These processes are fundamental building blocks for effective network management and security. Authentication provides a way of identifying a user, typically by having the user enter a valid user name and valid password before access is granted. The process of authentication is based on each user having a unique set of criteria for gaining access. Authorization is the process of finding out what an authenticated user can do within the system. The authorization process determines whether the user has the authority to perform such actions. Accounting is the process of logging the activity of an authenticated user, e.g., the amount of data a user has sent and/or received during a session, called APIs, etc. Many types of research were done in the traditional AAA. Especially, in the 5G/B5G network area, Zarca *et al.* [176] proposed the AAA management method

in NFV/SDN-enabled IoT scenarios. For the virtual network, Wong *et al.* [177] proposed a virtualized AAA (V-AAA) for the 5G network.

In NSAI, the traditional centralized AAA structure is facing great challenges because of the increasing demand from access devices and more frequent access authentication requests [178]. Meanwhile, the authorization and accounting tasks are also very increased in volume, due to larger scale applications. From the perspective of the GSS Tier, blockchain technology is providing a promising framework for distributed AAA, as illustrated in Fig. 13. As a decentralized data structure, the strong immutability, nonrepudiation, and auditability of blockchain contribute to the safer operation of the authentication system [179], and the pressure of data transmission and processing can be significantly relieved [180]. With the application of cryptographic algorithms, such as group signature and ring signature in the data structure, the data can also have higher anonymity [181]. Kiyomoto *et al.* [182] proposed a blockchain-based authorization architecture for B5G mobile services, for the problem of information disclosure caused by identity authentication. Kravitz and Cooper [183] proposed to use the licensed blockchain to protect and manage multiple IoT nodes. The system provides a distributed identity management solution, which improves security and prevents attacks by rotating asymmetric keys. Hashemi *et al.* [184] proposed to improve access management by defining a blockchain-based multilevel mechanism, which would specify capabilities, access lists, and access rights. Rashid and Pajohoh [185] proposed a security framework for IoT authentication and authorization based on blockchain technology. Pouraghily and Wolf [186] proposed a ticket-based verification protocol for blockchain transactions accounting on IoT devices. Franco *et al.* [187] introduced a blockchain-based reverse auction solution that can help to reduce costs involved in VNF's commercialization and monetize NFV-enabled infrastructures.

As a result, the NSAI architecture can be presenting a convergence of 5G/6G, AI, and blockchain. The blockchain could be implemented into two parts of NSAI. One is the ledger system using the Proof-of-Stakes (PoS) mechanism to manage the credit and trust in the AAA system. Another is the blockchain node integrated on the virtual service level to provide user-specific AAA service. The hierarchical and distributed V-AAA architecture for 5G systems [177] was conceived to handle multitenancy, multinetwork slicing, and multilevel services. A new hierarchical and distributed database architecture is also considered to interwork with V-AAA, capable of coping with the network flexibility, elasticity, and traffic fluctuation.

It is however worthwhile to note that the blockchain technology still needs to be further improved, in order to satisfy the requirements of NSAI AAA, especially in its throughput and latency, in order to satisfy the real-time massive data transactions in the system.

D. Application Tier

The role of the APP Tier is to provide users with specific and groundbreaking user experiences, based on smart services

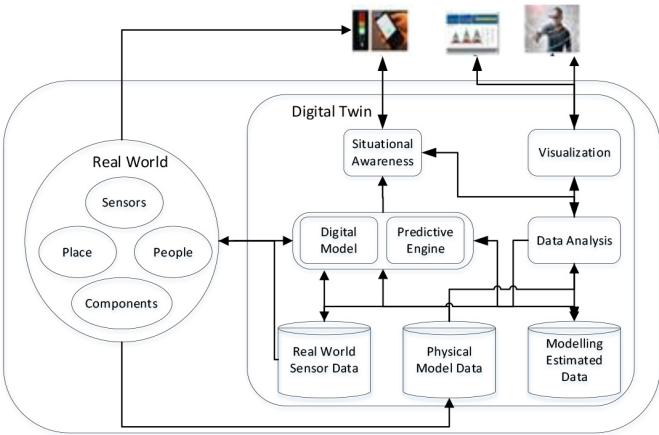


Fig. 14. Digital twin architecture [257].

provided by the bottom GSS tier. The APP Tier can realize intelligent planning and management, support automation service, smart city, smart industry, smart transportation, smart grid, smart health, and other intelligent applications. As GSS virtually transforms NSAI into a distributed computing system that converts massive real-time data into knowledge, the APP Tier can use technologies, such as digital twin, augmented reality (AR), virtual reality (VR), and knowledge graph (KG) to build variable applications. The digital twin can be utilized to reflect and connect the physical world to the digital world. AR and VR can be utilized to implement augmented and immersive experiences. KG [188], [189] can be utilized to implement knowledge reasoning, and understanding the information semantics, thus generating the optimal decision.

1) Digital Twin: Digital twin [190] is a digital replica of a living or nonliving physical entity which includes potential and actual physical assets (physical twin), processes, people, places, systems, and devices that can be used for various purposes [116]. As shown in Fig. 14, the digital representation provides both the elements and the dynamics of how a smart device operates and lives throughout its life cycle. Definitions of digital twin technology used in prior research emphasize two important characteristics [192]. First, it emphasizes the connection between the physical model and the corresponding virtual model or virtual counterpart [193]. Second, this connection is established by generating real-time sensor data [194]. The concept of the digital twin can be compared to other concepts, such as cross-reality environments or co-spaces and mirror models, which aim to synchronize part of the physical world (e.g., an object) with its cyber representation [195]. The digital twin can integrate IoT, AI, machine learning and software analytics with spatial network graphs [196] to create living digital simulation models that update and change as their physical counterparts change. Digital twin continuously learns and updates itself from multiple sources to represent its near real-time status, working condition, or position. The digital twin also integrates historical data to factor into its digital model [197].

The specific information contained in the digital twin is driven by use cases. An example of digital twin is the use

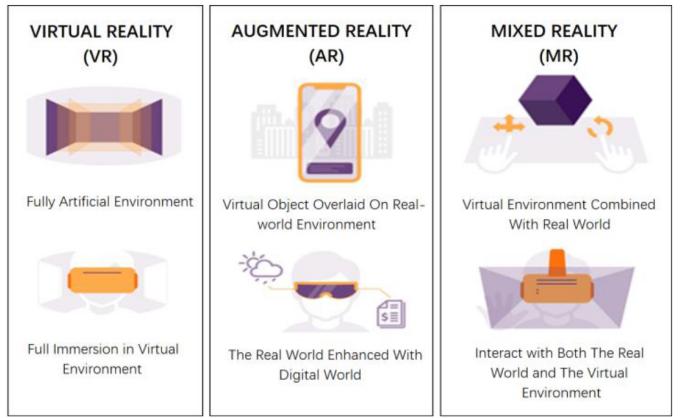


Fig. 15. Immersive technology [258].

of 3-D modeling to create digital companions for physical objects [198]. It can be used to view the status of the actual physical object, which provides a way to project physical objects into the digital world [199]. Digital twin originated in the field of industrial manufacturing, but it has been applied to various fields of smart city, including urban microspace, transportation, education, medical, and energy. Under the framework of NSAI, a more powerful digital city replica can be built. In this digital replica, the operating status of infrastructure (water, electricity, gas, transportation, etc.), the deployment of municipal resources (police, medical, fire, etc.), etc., will be collected through sensors, cameras, and digital subsystems, and delivered to the digital replica.

2) Immersive Computing: Immersive technology refers to technology that attempts to emulate a physical world through the means of a digital or simulated world by creating an immersive feeling. As illustrated in Fig. 15, immersive technology enables mixed reality (MR), which is a combination of VR and AR, or a combination of physical and cyberspaces [200]. In some cases, the term "immersive computing" is effectively synonymous with MR as a user interface [201], [202].

VR is a simulated experience that can be similar to, or completely different from the real world. State-of-the-art VR systems use either VR headsets [203]–[206] or multiprojected environments to generate realistic images, sounds, and other sensations that simulate a user's physical presence in a virtual environment [207]. A person using VR equipment can look around the artificial world, move around in it, and interact with virtual features or figures. The effect is commonly created by VR headsets consisting of a head-mounted display with a small screen in front of the eyes but can also be created through specially designed rooms with multiple large screens. VR typically incorporates auditory and video feedback, but it may also allow other types of sensory and force feedback through haptic technology [208].

AR [209] is an interactive experience of a real-world environment where the objects in the real world are enhanced by computer-generated perceptual information, sometimes across multiple sensory modalities, including visual, auditory, haptic, somatosensory, and olfactory [210], [211]. AR can be defined

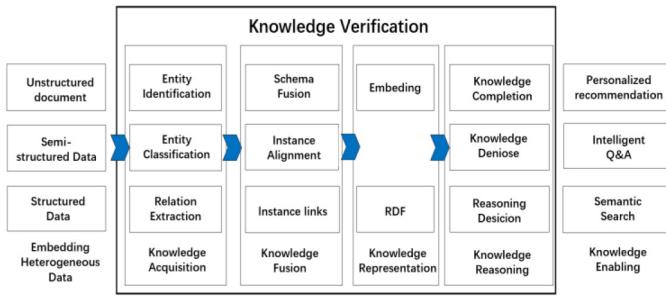


Fig. 16. KG technology chain.

as a system that fulfills three basic features: 1) a combination of real and virtual worlds; 2) real-time interaction; and 3) accurate 3-D registration of virtual and real objects [212]. As a result, AR alters one's ongoing perception of a real-world environment, whereas VR completely replaces the user's real-world environment with a simulated one [213]. AR is related to two largely synonymous terms: 1) MR and 2) computer-mediated reality. The primary value of AR is how components of the digital world are blended into a person's perception of the real world through the integration of immersive sensations. The information about the surrounding real world becomes interactive and can be digitally manipulated, which can be either virtual or real. Augmentation techniques are typically performed in real time and in semantic contexts with environmental elements.

Together with AR/VR and other technologies, immersive experiences can be provided in several areas, including retail and e-commerce [214], art and entertainment [215], military [216], education [217], and medicine [218]. It is also growing in the nonprofit industries, such as disaster relief and conservation, by putting a user in a situation that would elicit more real-world experience and giving them a stronger emotional connection to the situation. As immersive technologies are becoming mainstream, they will likely pervade in other industries as well.

3) Knowledge Graph: As shown in the technology chain of Fig. 16, knowledge acquisition refers to extracting knowledge of related entities, attributes, relationships, and events from different sources and structured data. The entity is the core unit of KG. The traditional entity recognition method is dominated by statistical models, such as HMM and CRF. BiLSTM + CRF [219] model avoided the feature template construction of traditional CRF, and two-way LSTM was utilized to make better use of the semantic information. Entity classification is the classification of extracted entities. Relationship extraction is the automatic extraction of specific semantic relationships between entities to supplement the missing relationships in the graph. The method based on bootstrap learning extracts new relationships with a few seed instances or templates and then generates more templates with new results, e.g., KnowItAll [220] and TextRunner [221]. Remote supervision [222] took existing triplet information as seed and matches information containing both subject and object as annotated data of relationships. To further improve the accuracy, Zheng *et al.* [223] proposed to use an end-to-end model of a neural network to realize entity recognition

and relation extraction at the same time, to avoid the cumulative impact of error accumulation caused by the results of early entity recognition on relation extraction.

Knowledge fusion mainly solves the problem of multi-source heterogeneous data integration. Entity fusion mainly involves schema fusion, entity alignment, entity link, and other technologies [224]. A schema is a model for a KG whose fusion is equivalent to a merge of the type and properties. Zhao *et al.* [225] comprehensively introduced the latest research progress of opensource knowledge fusion, multi-knowledge graphs fusion, information fusion within KGs, multimodal knowledge fusion and multisource knowledge collaborative reasoning. Xu *et al.* [226] introduced the knowledge fusion taxonomy to understand the relationships among traditional marketing analytics, big data analytics, and new product success. For instance, alignment, it can be regarded as a sorting problem to find the top matched instances, or a dichotomy problem to find whether to match. Its characteristics can be obtained based on entity attribute information, schema structured information, semantic information, etc. Entity alignment is an important process in multisource data fusion, where the data from different knowledge-based systems are distinguished for the same entity, and then fused to generate the unique entity in the knowledge base. Once a KG is constructed, how to accurately match the corresponding entity in the graph and then extend the relevant background knowledge is an entity link problem. Entity links primarily rely on a many-to-many mapping table between entities and need to correctly locate the entity mentioned by the user, understand the real expression intention of the user, to further explore user behavior, and understand user preference.

Knowledge representation is a description and agreement of knowledge data, making the computer understand knowledge like a person. Most KGs use a symbolized method for representation, where the resource description framework (RDF) is the most commonly used symbolic semantic representation model. For example, <Subject Subject, Predicate, Object Object> expresses an objective fact, and the method is intuitive and understandable, with interpretability and reasoning. The embedding algorithm based on vector presentation has emerged to train a representable vector for each entity and relationship, which can characterize and further explore hidden knowledge. Common embedding models include Word2Vec and Trans series [188], [227].

Reasoning based on KG aims to derive new knowledge based on existing knowledge information, including entity relationship, attribute, etc., or identifying error relationships. It can be divided into symbol-based reasoning and statistic-based reasoning. The former generally creates new rules of entity relations according to classical logic, or judges the contradictions of existing relations, while the latter learns new entity relations from the graph through statistical rules. For example, to prevent users from carrying out lateral movement attacks in an edge–cloud computing environment, Tian *et al.* [228] proposed a real-time lateral movement detection method, named CloudSEC, based on an evidence reasoning network for the edge–cloud environment. Wang *et al.* [229] proposed a scheme to integrate knowledge reasoning and semantic

data where the reasoning engine processes the ontology model with real-time semantic data from the production process. Jiang *et al.* [230] proposed a knowledge-bridge graph network (KBGN) model by using a graph to bridge the cross-modal semantic relations between vision and text knowledge in fine granularity, as well as retrieving the required knowledge via an adaptive information selection mode. Chen and Luo [231] proposed an automatic literature KG and reasoning network framework based on ontology and natural language processing, to facilitate efficient knowledge exploration from literature abstract.

A KG contains rich semantic information and a deep understanding of semantics. It can provide more direct and accurate query results in the fields of recommendation, search, and question answering. Through the relationship between entities, personalized recommendation extends the similar entities that users prefer, and provides interpretable recommendation content. On the one hand, the graph provides feature information of entities in multiple dimensions; on the other hand, the representation learning vector contains certain semantic information, which searches for recommended entities favorable to users' preferences. For example, Chen *et al.* [232] proposed to utilize opensource knowledge resources autonomously for human–robot interaction. Liu *et al.* [233] proposed a knowledge-enabled language representation model with KGs, where triples were injected into the sentences as domain knowledge.

IV. EXEMPLARY APPLICATIONS

We now present a few exemplary applications of NSAI, which typically require strong integration and interactions among cyberspace, the physical world, and target users. In terms of smart service provision, the requirements can include more real-time and reliable service rendering to ensure a strong interactive user experience and safety, and more intelligence in terms of AI agents learning from the open operating environment and human requirements in an online-evolutive fashion. Without the paradigm shift as presented by NSAI, the service reliability and target-user experience of these application scenarios cannot be possibly realized, especially in the dynamic networking and service environments. The features of “real-time intelligence” will also be an enabler for the next-generation Internet-based on NSAI.

A. Urban Microspace Service

Billions of devices and sensors are being deployed over global urban areas, which automatically collect data on everything from traffic to weather, energy usage, water consumption, shopping information, carbon dioxide levels, and more. The data are aggregated and transported to stakeholders where they are stored, organized and analyzed to understand what is happening and what is likely to happen in the future. NSAI-based urban microspace service connects and integrates the physical environment and the real-time events into an AI-based virtual system. With the popularity of digital-twin technologies, the system can converge all involved parties, online and offline, physical and digital world. Smart urban microspace

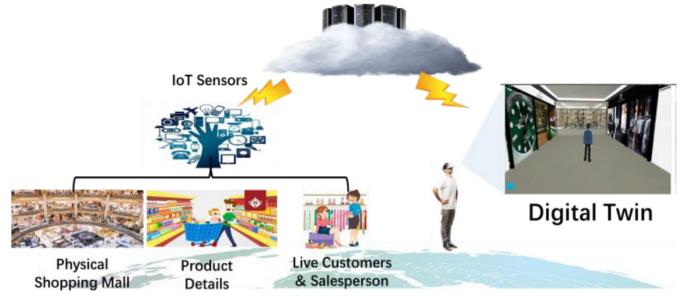


Fig. 17. Example of urban microspace service.

services can include smart community, smart building, smart work-space, smart shopping mall, other smart zones, etc. Some key features [234] of the urban microspace service can include the following.

1) *Convergence of IoT and Mobile Internet:* The future microspace service will converge sensor networks, IoT, Internet, and mobile networks, within the NSAI architecture and framework. All data can be integrated into the NSAI, and analyzed, mined, reused in real time. Meanwhile, all analyzed cyber data can be feedback to the corresponding physical world. The remote connecting and controlling of sensor devices can be more convenient. The number of links among the devices and the efficiency of data transmission are greatly improved, which can realize ultrareliable large-scale machine-to-machine and human-to-machine interactions. The Internet, 5G together with the IoT, will work as integral parts of the NSAI virtual networks and services, providing customized service based on the specific user requirements.

2) *Synchronization of Online/Offline Services:* All offline information, e.g., from the physical world, can be digitalized and synchronized with the online cyber system in real time. All instructions and events generated by the users or supervisors in the digital world shall also be fed back to the physical world. Based on the synchronization, people can get to know everything about the status, events, and services of the offline microspace by the digital online system. Digital twins of streets, communities, shopping malls, and buildings and commerce can be generated, making it easier to monitor and manage physical facilities online, and to search, view and use the services as provided by the physical world.

3) *Immersive User Experience:* Based on AR/VR technologies, one can build an identical digital copy of the physical world. The new digital world is in parallel to the corresponding physical world, and all actions made in the digital world can be realized in the physical world. Meanwhile, AI will play a key role to improve the user experience both in the digital and the physical worlds. The immersive experience allows users to experience, look around, and interact in the real-time 3-D space-time scene at any time and anywhere.

The NSAI-based urban microspace service can provide users with unprecedented experience beyond the current mobile Internet. For example, in Fig. 17, John suddenly feels like going to Lafayette in Paris for shopping. He does not need to suffer the long trip by flights, taxi and trains. He can just do it virtually with the smart shopping mall service. The physical

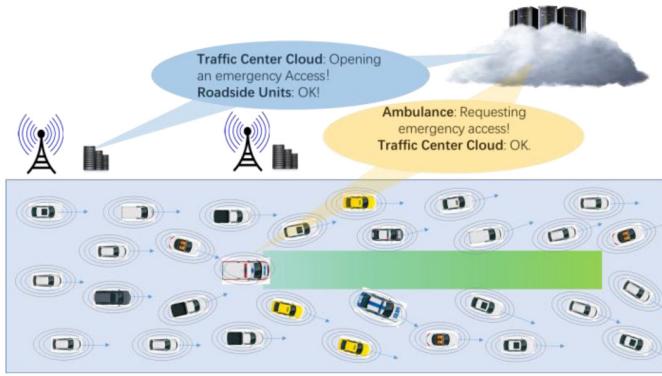


Fig. 18. Example of transportation service.

Lafayette shopping mall can be copied to the virtual world so that John can experience the real-time Lafayette online. Virtually, he can feel the physical world like weather, a new decoration for Christmas, customers and crowd, brands advertisement, promotion, in-stock products, etc. John is involved in the digital world as a digital customer and looks around or shops whatever he wants. He can walk into the LV or Channel shop to check the new arrivals. The item he selected will be shown to him in a 360-degree 3-D VR scene. He can talk to the clerks in real time to confirm the information, promotion, or price of the selected product. If he wants to buy, the product will be charged and then delivered to his home automatically. He can also use an AI chatbot to search the items in his wish list and get precise results in real time. The whole “go shopping in Paris” process can be implemented by the NSAI-based service seamlessly and smoothly, giving consumers a real-time and rich interaction experience.

B. Transportation Service

NSAI-based intelligent transportation systems converge people, vehicles and road infrastructure into a massive but integrated AI entity. It can collect vehicle and road information in real time, and control its operation, monitor intersections, management information, traffic conditions, vehicle speed, etc. With the service, people can experience new functions, such as a shared individual trip, customized group trip, fully autonomous driving, etc. Some key features [235], [236] can include the following.

1) Fully Autonomous Driving: NSAI can accelerate the transformation from the Internet of Vehicles to fully autonomous driving. Internet of vehicles is a large system with wireless communication and information exchange between vehicles and everything. Internet of vehicles can be the basis for fully autonomous driving, which presents an integration of AI, visual computing, radar, monitoring devices, and global positioning systems. Full autonomous driving not only relies on efficient and real-time data computation but also needs real-time access to road infrastructure and Internet data. As a result, fully autonomous driving is no longer limited by the computing and sensing capability of any single vehicle, but depends on the whole network to coordinate all vehicles and road infrastructure for efficiency and safety.

2) Vehicle-Road Synergy: The development of vehicle-road synergy (VRS) realizes intelligent collaboration between vehicles and road infrastructures. VRS can ensure traffic safety, improve traffic efficiency, reduce urban pollution, and, thus, form a safe, efficient, and environmental-friendly road traffic system. NSAI-based VRS can present the new trend of the intelligent transportation system, where the integration of road network, sensor network, control network, and energy network can accelerate the simultaneous operation of different levels of smart vehicles on the same road infrastructure.

3) Smart Traffic Management: The NSAI-based intelligent transportation system can realize the convergence of roadside stations, onboard equipment, smartphones, and other communication terminals, to set up a control center for urban traffic management with a global view. Smart traffic management can update the routes of all vehicles in real time to avoid urban traffic congestions. It can also provide emergency assistance to ambulances and police cars where the road can be automatically cleaned without halting the traffics.

The intelligent transportation service will provide people with groundbreaking experiences. For example, in Fig. 18, on a fully automated driving road, traditional traffic-light-based control will become history. All vehicles can automatically, orderly and safely stop, turn, change lanes, move back, and move forward under the collaborative work of vehicle, road, and cloud. John can read newspapers or prepare meeting materials in his self-driving car and can also change destinations at any time through the chatbot in the car. In emergent situations, such as an ambulance applying for an emergency passage, the NSAI-based traffic service replans all vehicle operating paths in the cloud and roadside servers, and provides the ambulance with a cleaned green lane without any traffics or obstacles in real time.

C. Healthcare Service

NSAI-based healthcare services integrate the doctors, patients, hospitals, health management agencies, and government supervisors as a networking system and provide seamless smart services to all involved parties. Healthcare service can be provided continuously with high quality through a life-long cycle. All the data generated by people, patients, doctors, hospitals staff, or equipment can be stored, secured, retrieved, analyzed, or shared. Medical resources can also be shared among doctors and hospitals. Patients can be checked by a doctor or medical equipment remotely, and doctors can also retrieve all the shared data from either local or remote databases. Therefore, the physical boundary of hospitals and the geographical boundary of cities would be eliminated in healthcare. The NSAI-accelerated medical service can improve the efficiency of doctors and hospitals significantly, and revolute state-of-the-art health management, treatment, and hospital management. Some key features [237], [238] of the NSAI-based healthcare service can include the following.

1) Connected Healthcare: The whole life cycle of healthcare service is being rebuilt, by connecting residents, doctors, hospitals, medicine, health management, medical care, etc., to provide healthcare-relevant services, including health



Fig. 19. Example of healthcare service.

assessment, doctor appointment, health/hospital files, birth certificates, vaccinations, home delivery of medicines, medical checking and inquiries, health status notification, and others. Doctors can retrieve the health data with authorization to accelerate the precise diagnosis. For special groups of patients like hypertension, diabetes, infants and young children, and pregnant women, their real-time status can be traced promptly.

2) Smart Patient Monitoring: Patient monitoring systems, including bedside monitors, require more flexibility, reliability, and security. By the NSAI-based healthcare network, patient monitors can roll alongside patients as they moved from one room to another, with continuous smart monitoring. The AI-based controller automatically recognizes whenever a patient monitoring device joins the network, where real-time data are processed, analyzed, and streamed by the networking system.

3) Telehealth Applications: Telehealth services can overcome geographic barriers, reduce the time spent between diagnosis and therapy, and help in the early identification of health issues, especially in many developing countries. Telehealth systems are composed of healthcare facilities with different complexities and practices. In most countries, primary care facilities are the gateway to patients in the healthcare network, where patients can be further referred to specialized care in clinics or hospitals. The diversity of telehealth modalities can be used in different contexts and specialties. More advanced telemedicine innovations require higher quality, uninterrupted network service. Telesurgery, for example, could allow for expertized surgeons to help local surgeons in performing emergency procedures from remote locations.

The NSAI-based healthcare service connects all medical resources and involved parties, to realize ubiquitous and personal-customized health management for everybody. The system adapts to everybody and continuously improve automatically. For example, in Fig. 19, John is a senior person who is under monitoring by the healthcare service. His health data can be obtained from wearable sensors daily and transmitted to the system in real time. If he falls accidentally, the system can detect the accident immediately and send an ambulance to pick him up to the proper hospital automatically. In the hospital, the doctor can retrieve all his healthcare and medical data to decide if he needs more diagnoses. If required, remote diagnosis can be applied by the equipment in another hospital. If John may need surgery, remote doctors may perform a



Fig. 20. Example of educational service.

telesurgery system when necessary. After the medical care, John will be further monitored by the smart patient-monitoring system seamlessly. After discharge, his health condition can still be analyzed continuously at home and in the community.

D. Educational Service

NSAI-based educational services similarly integrate the students, teachers, schools, academies, universities, libraries, and all learning resources, and provide seamless service to students, teachers, and supervisors. Life-long learning can be realized with the AI-based educational network and service. Besides human teachers, students can also learn from AI teachers. The data generated by students, teachers, publishers, or other parties can be stored, secured, retrieved, analyzed, or shared easily. The physical boundary of schools and the geographical boundary of cities would then be eliminated. The NSAI-accelerated educational service will improve the efficiency of learning and teaching significantly and solve the uneven learning-resource problem among cities and countries. Students can also benefit from the personalized and customized learning/training procedure, where everybody can be educated uniquely. Some key features [239] can include the following.

1) Adaptive Learning: The rapid development of AI and networks has made the so-called “adaptive learning” possible. First, a large amount of data about students’ learning behavior and preference can be generated in real time, where teaching methodologies and interactions can be dynamically reconfigured accordingly. The efficiency of the learning process can then be increased continuously. A student’s learning progress is not only determined by the teacher now but also by the machines that observe and analyze the student’s learning process in real time.

2) Smart Class: With AR/VR technologies, students can explore concepts by touching, pinching, and zooming the subject. Teachers can automatically log in to the smart-class system as soon as they enter the classroom, where students can deliver feedback digitally. Children with special needs may require more frequent or full-time assistance from teachers, where robot applications can help students continue their education outside the classroom, delivering the same responsiveness as in the classroom to their phones or pads, regardless of the distance or location.



Fig. 21. Example of industrial service.

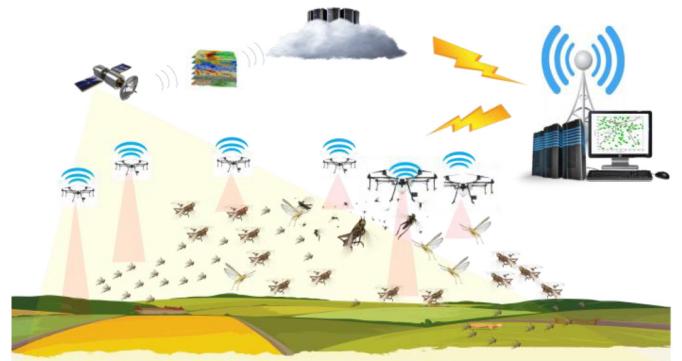


Fig. 22. Example of agricultural service.

3) *Virtual Classroom*: A virtual classroom is an online space that simulates a live classroom. Learning sessions are usually synchronous with teachers and students to interact in an online space in real time. State-of-the-art virtual classroom often includes the following features: video conferencing that facilitates communication; a digital whiteboard that offers real-time explanations and/or collaboration; participation controls that students can still “raise their hands” or otherwise participate in lessons; and subchats or group chats that students can collaborate in small groups online, e.g., breakout rooms. Virtual classrooms can range from an optional supplementary resource to the whole educational program in curriculum.

The NSAI-based educational service is changing the way of learning and teaching, which can intelligently adapt to every person and continuously improve. For example, in Fig. 20, John has a daughter Jennie in the age of primary school. He can choose to join her class remotely when needed. When Jennie is talking, the system can track her emotions, analyze her voice, and develop a real-time emotional state diagram. Therefore, the system can check whether she is paying attention and remind her accordingly. The system can also help the teacher to find if Jennie understands a specific concept. Her learning dynamics can be captured in real time so that the most suitable learning path for her can be customized and recommended to improve her learning efficiency. Personalized and customized learning can be realized. After the class, an AI teacher can continue to expand Jennie’s learning process and help to develop her interests and specialties.

E. Industrial Service

The industrial Internet is deemed as a key infrastructure linking the entire industrial system, industrial chain, and value chain to support the development of industrial intelligence. NSAI-based industrial services connect the customer order, manufacturing, logistics, sales, and warranty and integrate people, design houses, factories, warehouses, and supply chain agencies into one networking system. Some key features [240], [241] can include the following.

1) *Customer-Driven Industrial Model*: The driving force of all product iterations is to meet customers’ needs, discover value from customers, and transform it into a real business. With the NSAI-based industrial service, fast, small-batch, customized products to meet the individual needs of customers are

becoming a reality. The supply, manufacturing, and sales can be digitized, where rapid, effective, and personalized industrial products can be realized. Through the massive data analysis, the system can predict customer behaviors, to greatly improve user experiences; and it can also reduce the mismatch between production and sales at the same time.

2) *Smart Manufacturing Process*: From the perspective smart manufacturing process, the industrial Internet can be summarized with three key points. Data collection is the foundation where ubiquitous sensing technology implements efficient data collection and aggregation from multisource devices, heterogeneous systems, operating environment, personnel information, and other elements. The industrial service platform is the core that builds an extensible operating system on the existing mature infrastructure, to provide a basic platform for industrial application development. The industrial applications are the key for specific industrial scenarios in promoting the modeling and software packaging of industrial technology, experience, knowledge, and best practices.

3) *Convergence of Automation and Informatics*: Through NSAI, an interconnection can be formed between people and machines, machines and machines, and services and services, thereby achieving a high degree of horizontal, vertical, and end-to-end integration. Industrial smart service presents the convergence of automation and informatics which can create the end-to-end production process guided by the value chain, realizing the effective integration of the digital world and the physical world, and converging the product value chain with different companies and customer needs.

For example, in Fig. 21, John has just renovated his house and needs to order a set of new furniture. He places an order through the furniture factory’s system, which can be an entry point of the following smart manufacturing services. First, John can easily reconstruct the house in a 3-D digital model by his smartphone, and an AI-based designer can interact with him on the customized design, including styles, colors, materials, etc. After John confirms his design scheme, production, and supply chain are fully customized and automated based on John’s design. The manufactured furniture is directly delivered to John’s house through the smart logistic system. The entire process can be carried out without manual intervention, and all states and procedures can be controlled and tracked in real time, providing highly accountable smart manufacturing.

F. Agricultural Service

According to predictions from United Nations, the global population will reach nine billion by 2030. More efficient agricultural production will be the only way to feed the world's growing population. Traditional agricultural production mainly relies on experience, which not only may waste human and material resources but also can damage the Earth's environment. The circulation of traditional agricultural products makes it difficult to accurately predict the market requirement, weather, pests, and diseases. NSAI-based agriculture services can be an integration of emerging technologies, to achieve precision farming, visual management, and intelligent decision making in agricultural production. Toward this end, it connects all involved parties, including consumers, agriculture experts, farmers, and dealers together to improve the efficiency and precision of agricultural production and circulation. Some key features [242], [243] can include the following.

1) *Agricultural Environment Control*: Agricultural environment monitoring and prediction is the basis of future agriculture service and can accelerate agricultural production efficiency. By the real-time data analysis technology over NSAI, the growth of crops, etc., can be predicted; the moisture content, seedling condition, insect condition, and disaster situations can be forewarned and analyzed; and environment pollution information can be closely monitored.

2) *Unmanned Field Operation*: Unmanned agricultural machinery, such as harvesters and tractors can work around the clock. Moreover, unmanned agricultural machinery can automatically and precisely plan routes, turn around, collaboratively operate, and efficiently complete tasks. High-definition video of field operations can be smoothly transmitted back to the remote-control center in real time. Agricultural machinery can also explore the surrounding environment in all directions at high speed.

3) *Precision Farming*: The smart irrigation system uses high-precision soil temperature and humidity sensors and smart weather stations to remotely collect data, such as soil moisture, PH value, nutrients, meteorological information, etc., to realize automatic moisture or drought forecasts, intelligent decision making on irrigation water consumption, remote and automatic control of irrigation equipment, etc. With precision farming, accurate fertilization and proper irrigation can be achieved. Meanwhile, smart irrigation can also automatically adjust the required nutrient solution according to a specific plant growth rate. Smart livestock accurately monitors the live-stock reproductive process and growth dynamics, to realize fast and efficient guidance for livestock breeding and building management. Epidemic prevention can also be implemented.

For example, in Fig. 22, John's farm uses the smart agricultural system to manage the whole production process automatically and precisely. The experts and expert database can support the process as required. When a locust plague occurs, satellite remote sensing can detect it in real time. After computing and analyzing the disaster data through NSAI, route plan and pesticide spraying pattern for a drone cluster can be made automatically, and locust plague real-time images can be sent back by the drones. Through the NSAI-based service



Fig. 23. Example of energy service.

through satellite sensing, IoT, drones cluster AI planning, and intelligent networking, the locust plague can be eliminated automatically and efficiently, to minimize potential damage.

G. Energy Service

Clean energy has been the direction of future energy development. NSAI smart-energy service can realize not only multienergy complementation, such as wind power and solar energy but also the integration of power grids, heat grids, fuel grids, and their transportation networks. The energy Internet is becoming a wider and stronger smart grid with an ultrahigh-voltage power grid as the backbone, the transmission of variable clean energy as the lead, and interconnection and ubiquity as key features. It combines the energy industry with information technology and integrates energy production, storage, transmission, and consumption together in managing the entire energy ecosystem. Some key features [244], [245] can include the following.

1) *Convergence of Traditional and Clean Energy*: Clean energy power stations, such as solar energy and wind power, are making the power supply and power consumption mode more diverse and volatile. Traditional and renewable energy will be integrated and connected as a whole system. NSAI-based energy services can change the traditional centralized AI into a large distributed intelligent network to realize self-organizing, self-evolving, and real-time intelligence. The deterministic low-latency and the flexible network-resource configuration can provide a guarantee for "more real-time" system response; and distributed AI can provide a more accurate and online-trainable model with real-time data, to achieve "smarter" scheduling and control.

2) *Active Distribution Networks*: The intermittent and volatile characteristics of clean-energy power generation bring difficulties to the grid power balance and operation control. The distribution network has changed from a passive network with unidirectional power flow to an active network with bi-directional power flow, which requires new monitoring and communication systems to achieve the goal of reliability, flexibility and efficiency. NSAI can effectively resolve the problems of distributed power plants, scattered points, and large quantities of clean-energy distributed power stations. The current distribution network has a wide variety of sensors, multiple data sources, and strong real-time

performance for monitoring objects, such as power distribution rooms/stations, overhead lines, switching stations, etc. The integrated networking system can therefore fully reflect the characteristics of a more real-time and smarter system.

3) Efficient Dispatch and Control of Smart Grid: Smart energy services can respond to energy inefficiency or energy failure, and quickly resolve the issue. It adapts to all energy production, storage, transmission, and sales procedure, to strengthen energy management, reduce operation and maintenance costs, and save energy as much as possible. It can also eliminate human intervention with decentralized real-time records, trace energy production and consumption, and help various users to intuitively understand the quantity and price of energy with the most applicable energy solution.

The NSAI-based energy service will intelligently improve the efficiency of energy usage and reduce energy waste, as shown in Fig. 23. On the demand side, smart meters for consumers like John or his company, and sensors along transmission lines can continuously monitor demand and supply. Equipment named “synchronized phase” can measure the grid power flow in real time, enabling operators to actively manage and avoid interruptions. These sensors communicate with the grid and modify the way of electricity consumption during off-peak hours, which can further reduce the workload of the grid and reduce the cost. On the supply side, NSAI will allow for a transition to increase the production of clean energy and minimize the natural intermittent interference, e.g., caused by changes in sunlight and wind. When the operation of clean energy exceeds a certain threshold, e.g., due to increased wind or sunny days, the grid will reduce the production of fossil fuels, thereby limit greenhouse gas emissions. All energy can be used as efficiently as possible, and fossil-fuel power generation becomes only required when necessary.

V. CHALLENGES AND RESEARCH OPPORTUNITIES

A. Achievable Capacity of Massive Wireless Networks

Shannon's theory on channel capacity has been the foundation of the information theory, which proved the throughput limit of a point-to-point communication link with unlimited coding latency. Shannon gave a simple equation to describe the channel capacity as $C = B \cdot \log(1 + \text{SNR})$, where B is the channel bandwidth, and SNR denotes the signal-to-noise ratio. What remains open in the point-to-point link is the achievable capacity subject to QoS constraints, e.g., latency. Moreover, in the network information theory, it has been only proved that the end-to-end capacity can be further increased with the number of nodes N , e.g., on the order of $O(B \cdot \sqrt{N})$ for static networks, and $O(B \cdot N)$ for dynamic and mobile networks. It remains unclear what is the achievable network capacity under end-to-end QoS constraints, e.g., flow throughput, latency, and energy efficiency.

It has been reviewed in this article that by introducing L2 cognitive multihop communications, a breakthrough on classical Shannon capacity can be expected especially for massive wireless networks. The latest evolution of cellular standards has already provided the architectures, such as NR-IAB and NR-Sidelink, where further technical contributions

are possible to be introduced on L2 multihop wireless communications. In massive wireless networks, power and rate control of the radio nodes can require further research investigations, since the combination of them not only changes all single-hop communication QoS parameters but also the single-hop radio range as well. Advanced radio power and rate control, e.g., by AI, will contribute to identifying the actual network capacity under end-to-end QoS constraints. A lot of groundbreaking research opportunities can still exist in this area, as the communication research community has been looking to break the limit of Shannon capacity over decades. It is also interesting to investigate the converged research with the latest development of massive MIMO and NOMA, which further differentiate the radio signals in space-time-frequency domains.

B. Intent-Oriented and Real-Time Network Configurability

The intent-oriented network presents a new shift of the SDN/NFV paradigm, which exposes more and more network configuration details to the supported smart services or applications. With more intelligence on network configuration from service and application perspectives, real-time network configurability becomes a real challenge and opens new research areas to be studied. Traditionally, network configuration takes place on a relatively large time scale. For example, in current IPs and operations, network configuration usually takes place when new devices join the network, and such configuration usually remains static until the device leaves the network or a specific failure happens. As a result, in the legacy study of network configurability, little research has been given on the real-time features of configurability, i.e., how swift a network can be responding to the configuration commands. In the target NSAI services, network reconfiguration can happen in a much smaller time scale, e.g., in terms of seconds or even milliseconds, to better support the dynamic changes of NSAI QoS. Such configuration can include network topology, resource allocation, routing and load balancing, security, and every other aspect of the SCN.

C. Soft and Multidimensional Resource Allocation

Due to the dynamic nature of network resource requirements in the SCN Tier, such as for computing, communication, and caching resources, higher resource utilization efficiency can be achieved, by having multiple virtual networks to share the same physical resource, i.e., soft resource allocation. In contradiction to soft resource allocation, hard resource allocation provides guaranteed physical resources to a certain virtual network. In traditional communication engineering, spectrum resource sharing has been well studied particularly in the media access control sublayer. However, when multidimensional and heterogenous network resources are considered, resource sharing over multiple virtual networks can require new methodologies and algorithms to be developed. This challenge also presents research opportunities where little previous work has been delivered. For the soft resource allocation in SCN, it can be envisioned that various network resource access mechanisms shall be developed, e.g., on a

priority or contention basis, for different combinations of heterogeneous resources. Soft resource allocation would be used along with hard resource allocation in SCN, to strike a fine balance between efficiency and performance.

D. Dynamic Networking Protocols for Distributed AI

As is well known, current IPs have been insufficient in handling network dynamics, such as frequent network topology changes or resource variations. Traditionally, to deal with device mobility, wireless communications have been limited to the very last hop, and mobile IP was developed to provide mobile device roaming, by setting up a home agent and a foreign agent in IP routing. However, mobile IP cannot deal with network topology changes, and even frequent user/device changes in the network can introduce unaffordable overheads. Under the dynamic changes in the SCN Tier, new networking protocols, e.g., for routing, transport, and load balancing, shall be developed to support the distributed AI processing on top. This challenge also presents great research opportunities to develop a whole new intelligent Internet under NSAI. Since the dynamic networking protocols can be constructed on top of L2 in the virtual networks, backward compatibilities to the current IP stack may be provided by overlay and underlay networking techniques.

E. Online Evolutive Learning-Based AI System

As we have reviewed previously in this article, multiagent learning, such as MARL and MDRL, can be presenting a foundation of the “OEL” in NSAI. State-of-the-art machine learning methods (supervised or semisupervised learning) require expensive human data labeling that is usually infeasible in NSAI applications. Moreover, the statistical data attributes in NSAI can also be nonstationary in most real-world applications. OEL can present the paradigm shift, where the NSAI shall be capable of adapting to the dynamic and nonstable environment changes, i.e., newly generated and real-time data, without human labeling or intervention. In other words, NSAI shall be able to generate new knowledge by exchanging information among the AI entities. For MDRL, particularly, great challenges still exist to provide convergence of the multiagent learning, and real-time interactions. On the other hand, many research opportunities reside in investigating the online-evolutive and multiagent learning in NSAI, e.g., the generalization and convergence, to provide a generalized framework of distributed AI computing in the GSS Tier.

Within the prospects of OEL, the divisions of machine learning, including supervised learning, unsupervised learning, and reinforcement learning could be united in a generalized architecture, where the smart environment can be providing real-time feedbacks to all the agents for online optimization. The smart environment is created by a deep integration of multiagents (i.e., the cyber world) and the physical world, where laws of physical world can enforce constraints for the convergence of the cyber world (i.e., the multiagent AI). It can be foreseen that all fundamentals of state-of-the-art AI algorithms may be revamped.

F. Multitime-Scale System Configuration

As discussed previously, traditional networking system configuration based on IPs takes place at the hourly or daily time scales. With the customized virtual network for smart services, the AI-empowered network configuration however can take places at much finer time scales, e.g., a second or even millisecond levels. As configuration itself introduces network overheads, it remains open to decide what configurations can be tuned at smaller time scales, and what configurations shall be reserved for larger time scales. For example, policies, topology, and network-slice management intuitively shall be done at larger time scales, whereas routing and load balancing shall be performed at smaller time scales. On the network side, with the toolsets of soft/hard multidimensional resource allocation, resource reconfiguration can take place at all time scales. On the application side, AI computation can often have a soft and elastic requirement as well, which can be dynamically configured, such as inference accuracy and latency. As both the network conditions and application requirements can be soft and dynamically changing, the GSS Tier shall be provided with a set of APIs for AI-empowered network and application configuration and joint optimization, while the technical details of such configuration APIs remain open for research.

G. Super Power Efficiency and AI Society

Power efficiency has been presenting great challenges to state-of-the-art AI computing. Due to the complexity of advanced AI models, it has been argued that the current training energy cost on some GPUs can be even higher than the actual energy cost of flying an airplane. Since the birth of AI, people have been looking to design an AI computer with similar functions or a similar number of operating neurons as a human brain. However, although current micro and nanoelectronics are being improved for 10–100 times more power-efficient over time, e.g., by adopting new IC and algorithm architectures as reviewed previously, state-of-the-art electronics still consume about one million times more energy than the human brain. With the development of NSAI, we can envision an “AI society,” where lightweight algorithms and computing devices can be possible collaboratively execute sophisticated AI tasks, by leveraging the growing communication capabilities. After all, we may have to consider the following fundamental differences between humans and computers. Although computers cannot be comparable to humans in training or executing neural networks so far, the rate of information exchange among computers far exceeds any human languages. How to realize the superpower efficiency in NSAI by integrating AI computing and communications, remains open to be further investigated.

H. Hardware and Software Implementation

The recent advances in microcontrollers have made it possible to execute sophisticated tasks that require high computational capabilities as well as memory usage, e.g., in the SHCity applications. Recent advances in communication networks have made it possible to transfer a large amount of data in real time. However, the requirements on terminal

devices remain very strict in terms of energy consumption and device size, and the cost of such devices still needs to be further reduced. In the new era of NSAI, several vendors are building their ecosystem where their devices may work or interconnect smoothly without any challenge. However, the interconnection between devices from different vendors and different systems remains to be a serious challenge. A universal middleware in the GSS/SCN Tiers can overcome such challenges, and potentially provide a standardized language for applications across different device vendors and network operators. These challenges in hardware (HW) and software (SW) implementation call for an open NSAI package for both HW and SW.

I. Security, Accountability, and Privacy

As reviewed previously, security requirements can be populated in all the tiers and layers, e.g., in physical communications, networks, services, and applications. By fully integrating AI with the network, more intelligent security mechanisms can be envisioned for NSAI, which include but are not limited to online risk analysis, trust management, and active defense. How to design security mechanisms in the individual tiers of NSAI remains open for research. Accountability issues are closely related to security but focused on validating the data and behaviors of all NSAI elements. Blockchain technologies have been advocated for system accountability, due to their fully distributed nature and auditability. But many open research problems still exist for fully integrating blockchain with NSAI, especially due to the massive volume of data generated by the networking system. Especially, state-of-the-art blockchain throughput and latency need to be largely improved, for logging and tracking the data and behaviors of all system users and infrastructures. New and scalable blockchain architecture needs to be developed to handle the massive users and data in providing global smart services.

It has been well argued that higher user privacy can be realized by conserving user data locally, e.g., in federated learning. However, under the architecture of NSAI and with the popularity of envisioned NSAI smart services, AI will be exposed to every aspect of people's daily life, far exceeding the extent of the current mobile Internet. User cognitions and behaviors, model parameters, e.g., beyond raw data, are private to the users, but are obviously exposed to the system. In order to deal with the privacy protection issues of the future smart society, as enabled by NSAI, it presents cross-disciplinary challenges from both technical and social research perspectives. On the one hand, privacy protection policies and legislation would be needed in human society to decide the authorization and access of user private information; and on the other hand, technologies need to ensure that every access and use of user private information can be authorized, traceable, and auditable. As the developments of computers, the Internet, smartphones, and mobile Internet have been continuously reshaping the human society over the past four decades, it is undoubtedly that NSAI will once again profoundly reshape the modern human society. It will be the responsibility of all research communities to

work together, in order to have humans and AI in an integrated community that meets the common interests of all mankind.

VI. ROADMAP TO UBN

The prospect of UBNs can be envisioned-based upon the NSAI framework but present an evolution of NSAI which incorporates human intelligence by ABCI. In particular, we consider that UBN can share the architecture of NSAI with human-intelligence agents, which can eventually converge human and AI, and seamlessly integrate cyberspace, the physical world, and human society. In presenting the roadmap to UBN, we hereby review stat-of-the-art research and propose a few generalized open research problems.

A. AI-Empowered Brain–Computer Interface

Brain–computer interface (BCI) research has been around since the 1970s. Traditional BCI uses electroencephalogram (EEG) to remotely control external devices for various purposes. Recently, BCI research can also involve human hearing and visual sensing, Kosmyna *et al.* [246] proposed to use the visual image as a brain-machine interface control strategy based on EEG. Min *et al.* [247] proposed processing for SSVEP-controlled BMI. However, the EEG signal of the cerebral cortex is usually very prone to outside noise during the signal acquisition process. Although aftermath noise reduction could be performed by advanced signal processing [248], relevant intervention cannot be carried out in real time. The development of fMRI provided alternatives to further understanding of the sensing cortex, for example, more than a dozen cortical regions have been found and identified to participate in visual functions [249], [250]. Some existing studies have introduced pathway models and then used fMRI to divide these functional areas into dorsal and ventral processing pathways by stereoscopic projection [251], [252]. As an invasive way to acquire brain signals, Xie *et al.* [253] classified several gestures with clear intraoperative craniotomy and electrocorticography (ECOG) signals during epileptic seizure monitoring and studied the location of BCI implanted into the cortex accurately. Hochberg *et al.* [254] implanted an electrode chip into the motor cortex of a paralyzed patient, enabling him to control the movements of household appliances, such as an electrical television and prosthetic limbs. Invasive brain signal acquisition has the advantage of being accurate and almost interference-free, as compared to EEG; it can also be specific to functional brain parts, and real time, as compared to fMRI.

Because of recent research, ABCI becomes an emerging research area differing from traditional BCI, with emphasis on human sensing reconstruction and brain–computer networking. Particularly, ABCI not only studies the reconstruction of human sensing from modeling real-time brain signals but also aims to provide a closed-loop controlling by stimulating artificial human sensing, *vice versa*. Some advances in invasive brain signal acquisition may provide promising futures of ABCI, by the accurate electrode implanting in functional areas of brain sensing, e.g., Neurolink [255]. Beyond the futuristic applications of ABCI on all consumers, current applications of

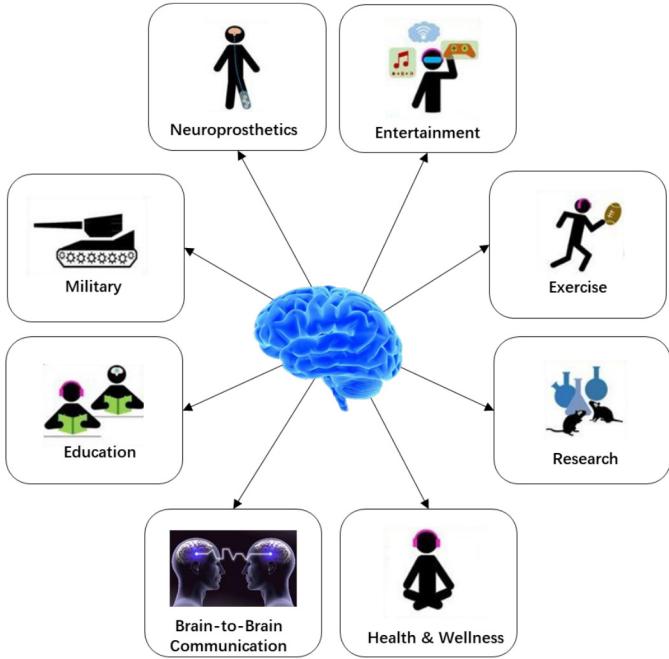


Fig. 24. UBNs.

ABCI include many areas of medical rehabilitation. For example, patients with cerebral palsy can control the mechanical arm by ABCI and gradually recover their movement abilities; and patients with vision, hearing, or tactile impairments can recover their sensations through ABCI.

B. Ubiquitous Brain Networks

Brain networks was interpreted as a network of Brain to Brain [256]. It allows everyone's brain to network together to solve complex problems and connect them to build a “social network” achieve “telepathic” communication. With the development of NSAI and ABCI, the UBN can be interpreted as a futuristic NSAI integrated with the human interface of ABCI, as illustrated in Fig. 24. In such a way, human intelligence will be a seamless and integral part of the NSAI, where human and AI will merge as one in UBN. Under UBN, everyone can connect to everyone by “telepathy,” and people can assign mind tasks to AI when thinking on the fly. When authorized, machines and robots can take human representation to communicate and explore new worlds in the Earth and the Universe. Human society will be united by mutual understanding and evolve with the explosive growth of intelligence.

UBN is more than a cross-disciplinary research of computing, communications, biomedical engineering, brain science, and psychology. Many open research opportunities exist since our current understanding of brain functions is still very limited. With the future migration from 5G to 6G, the roadmap to UBN is how NSAI technologies can embrace ABCI to converge human and AI. The following nonexclusive list of points are some of our considerations on the roadmap to UBN.

- 1) It is to determine which part of brain functions can be modeled and simulated by AI. Although current ABCI studies are investigating visual, hearing, language, tactile

sensing, etc., there can be other missing parts to give people a holistic virtual experience of the real world.

- 2) In assigning mind tasks to AI, it remains open *how accurate the brain can be interfaced with AI*. Although controlling, for example, wheelchairs or even drones have been realized, it is unclear if more complicated tasks such as operating a smartphone can be accurately captured. As human languages, e.g., English, are often limited in presenting the high-level abstract concept, it is unknown if UBN can more accurately interpret abstract concepts than any existing human language.
- 3) As initial applications are healthcare and rehabilitation related, it is still unknown *if UBN can help to relieve neural diseases*, such as Alzheimer’s disease, depression, and neural impairments introduced by cerebral infarction. It is however clear that AI can be an integral part of the brain operation to overcome such diseases or impairments.
- 4) It is still unknown *how UBN can revolute human intelligence*, by fundamentally changing the current human educational system, directly and accurately passing knowledge or experience into the human brain. It is however clear that new technology achievements will contribute to the explosive growth of human intelligence as an integral part of a larger UBN.
- 5) Beyond the obvious requirements on higher electronic and electrode advances, UBN will have a profound impact on the global human society beyond the Internet. For example, legislation on new privacy protection shall be required so that “mind” or “brain” data can be protected, and any access to the human brain or mind can be regulated and traceable by the UBN.

VII. CONCLUSION

NSAI contributes to the future society of online-evolutive AI, which can transform the current fragmented IoT industry, realize the deep integration with mobile Internet, and promote the next global IT revolution. However, the research of AI and communications have been performed in separate research communities, which can be limiting the pace of convergence in computing and communications. In this article, we have made our efforts in reviewing the state-of-the-art technical areas of the NSAI, by proposing a converged NSAI architecture, and reviewing the key technologies and application scenarios. From the communication perspective, NSAI takes the advantage of the evolution from 5G to 6G, particularly in the PN and SCN Tiers, where a UAI can be realized for integrated networks over EMBB, mMTC, and URLLC, and service-customized virtual networks can be created, configured, and dynamically reconfigured for every smart service provision. From the computing perspective, NSAI takes the advantage of distributed AI, for example, in the GSS and Application Tiers, where AI computing is becoming immersive in every element of the networking system. The whole networking system becomes a ubiquitous and online-evolutive AI computing machine that can be interacting with humans/users. With the development of ABCI technologies, we further envision the

roadmap to the UBNs, where the deep convergence of computing and communications can indeed change human society profoundly, with the seamless integration of human and AI.

Future work under the proposed NSAI framework should include case studies of different NSAI application scenarios where common approaches can be studied and obtained to support real-time intelligent services through joint optimization of communication networks and applications/services. In concluding this article, we also wish the work in this article can contribute to realizing a common research platform for the cross-disciplinary discussions in the converged areas of computing and communications; and help to promote a converged community for the open research, standardization, and industrial developments of the global NSAI.

REFERENCES

- [1] “GitHub—Karlrupp/Microprocessor-Trend-Data: Data Repository for My Blog Series on Microprocessor Trend Data.” [Online]. Available: <https://github.com/karlrupp/microprocessor-trend-data> (Accessed: Jun. 13, 2021).
- [2] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
- [3] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, “Deep learning for visual understanding: A review,” *Neurocomputing*, vol. 187, pp. 27–48, Apr. 2016.
- [4] T. Young, D. Hazarika, S. Poria, and E. Cambria, “Recent trends in deep learning based natural language processing,” *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018.
- [5] K. Ota, M. S. Dao, V. Mezaris, and F. G. B. D. Natale, “Deep learning for mobile multimedia: A survey,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 13, no. 3, pp. 1–22, Aug. 2017.
- [6] G. Litjens *et al.*, “A survey on deep learning in medical image analysis,” *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.
- [7] Y. Chen, Y. Li, R. Narayan, A. Subramanian, and X. Xie, “Gene expression inference with deep learning,” *Bioinformatics*, vol. 32, no. 12, pp. 1832–1839, Jun. 2016.
- [8] R. Li *et al.*, “Intelligent 5G: When cellular networks meet artificial intelligence,” *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 175–183, Oct. 2017.
- [9] R. Shafin, L. Liu, V. Chandrasekhar, H. Chen, J. Reed, and J. C. Zhang, “Artificial intelligence-enabled cellular networks: A critical path to beyond-5G and 6G,” *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 212–217, Apr. 2020.
- [10] S. Xu *et al.*, “RJCC: Reinforcement-learning-based joint communicational-and-computational resource allocation mechanism for smart city IoT,” *IEEE Internet Things. J.*, vol. 7, no. 9, pp. 8059–8076, Sep. 2020.
- [11] J. B. Bernabe, J. L. Canovas, J. L. Hernandez-Ramos, R. T. Moreno, and A. Skarmeta, “Privacy-preserving solutions for blockchain: Review and challenges,” *IEEE Access*, vol. 7, pp. 164908–164940, 2019.
- [12] A. Ghosh, V. Kulharia, V. Namboodiri, P. H. S. Torr, and P. K. Dokania, “Multi-agent diverse generative adversarial networks,” in *Proc. IEEE CVPR*, Salt Lake City, USA, Jun. 2018, pp. 8513–8521.
- [13] D. Li, X. Zhu, and L. Song, “Mutual match for semi-supervised online evolutive learning,” *Appl. Intell.*, to be published.
- [14] “Artificial Intelligence is Evolving All by Itself | Science | AAAS.” [Online]. Available: <https://www.science.org/content/article/artificial-intelligence-evolving-all-itself> (Accessed: Aug. 18, 2021).
- [15] J. Chen, K. Li, Q. Deng, K. Li, and P. S. Yu, “Distributed deep learning model for intelligent video surveillance systems with edge computing,” *IEEE Trans. Ind. Informat.*, early access, Apr. 4, 2019, doi: [10.1109/TII.2019.2909473](https://doi.org/10.1109/TII.2019.2909473).
- [16] S. Troia *et al.*, “Machine learning-assisted planning and provisioning for SDN/NFV-enabled metropolitan networks,” in *Proc. EuCNC*, Valencia, Spain, Jun. 2019, pp. 438–442.
- [17] V. Nguyen, A. Brunstrom, K. Grinnemo, and J. Taheri, “SDN/NFV-based mobile packet core network architectures: A survey,” *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1567–1602, 3rd Quart., 2017.
- [18] S. Chen, S. Sun, and S. Kang, “System integration of terrestrial mobile communication and satellite communication—The trends, challenges and key technologies in B5G and 6G,” *China Commun.*, vol. 17, no. 12, pp. 156–171, Dec. 2020.
- [19] “Qualcomm Addresses Growing Demand for 5G by Announcing New Snapdragon 768G Mobile Platform.” [Online]. Available: <https://www.qualcomm.com/news/releases/2020/05/10/qualcomm-addresses-growing-demand-5g-announcing-new-snapdragon-768g-mobile> (Accessed: May 16, 2021).
- [20] “ZTE Launches White Paper on AI Enables Network Intelligence—ZTE Press Release.” [Online]. Available: <https://www.zte.com.cn/global/about/news/2-6-2.html> (Accessed: Jun. 26, 2021).
- [21] “Ascend Developer Community-HUAWEI CLOUD.” [Online]. Available: <https://www.hiascend.com/en/ecosystem/industry> (Accessed: Dec. 28, 2021).
- [22] “Nokia Launches AVA 5G Cognitive Operations to Help Telcos Enter the 5G Era | Nokia.” [Online]. Available: <https://www.nokia.com/about-us/news/releases/2020/03/31/nokia-launches-ava-5g-cognitive-operations-to-help-telcos-enter-the-5g-era/> (Accessed: Oct. 17, 2021).
- [23] “Network Services Portfolio.” [Online]. Available: <https://www.ericsson.com/en/network-services> (Accessed: Apr. 2, 2022).
- [24] “How Orange Puts AI to Good Use in its Network Operations Pt. 1, AI & ML | TelecomTV.” [Online]. Available: <https://www.telecomtv.com/content/ai-ml/how-orange-puts-ai-to-good-use-in-its-network-operations-pt-1-38747> (Accessed: Dec. 15, 2021).
- [25] “Intel Demonstrates Nervana NNP for Cloud, Movidius Myriad VPU for Edge—CTIMES News.” [Online]. Available: <https://en.ctimes.com.tw/DispNews.asp?O=HK3BD4GKXHASAA00N8> (Accessed: Jun. 19, 2021).
- [26] “GSMA | Case Study: China Telecom—Future Networks.” [Online]. Available: <https://www.gsma.com/futurenetworks/wiki/case-study-china-telecom-2/> (Accessed: Dec. 24, 2021).
- [27] “Google Cloud Teams With AT&T to Offer New 5G Solutions.” [Online]. Available: <https://www.theburnin.com/technology/google-cloud-att-5g-product-development-partnership-2020-3> (Accessed: Dec. 27, 2021).
- [28] “Vodafone Launches Artificial Intelligence Framework.” [Online]. Available: <https://www.vodafone.com/news/public-policy/vodafone-launches-artificial-intelligence-framework> (Accessed: Dec. 28, 2021).
- [29] “T-Mobile U.S., Inc—From AI to VR and Beyond: T-Mobile Accelerator Names Class of 2020 Startups.” [Online]. Available: <https://www.businesswire.com/news/home/20200519005603/en/From-AI-to-VR-and-Beyond-T-Mobile-Accelerator-Names-Class-of-2020-Startups> (Accessed: Oct. 28, 2021).
- [30] “Haier, China Mobile and Huawei Launch the World’s First AI + 5G Interconnected Factory—Huawei.” [Online]. Available: <https://www.huawei.com/en/news/2019/7/first-global-intelligence-5g-interconnected-factory> (Accessed: Jul. 9, 2021).
- [31] H. Yao, T. Mai, X. Xu, P. Zhang, M. Li, and Y. Liu, “NetworkAI: An intelligent network architecture for self-learning control strategies in software defined networks,” *IEEE Internet Things J.*, vol. 5, no. 6, pp. 4319–4327, Dec. 2018.
- [32] Y. Dai, D. Xu, S. Maharjan, G. Qiao, and Y. Zhang, “Artificial intelligence empowered edge computing and caching for Internet of Vehicles,” *IEEE Wireless Commun.*, vol. 26, no. 3, pp. 12–18, Jun. 2019.
- [33] Y. Zhao, B. Yan, Z. Li, W. Wang, Y. Wang, and J. Zhang, “Coordination between control layer AI and on-board AI in optical transport networks [Invited],” *J. Opt. Commun. Netw.*, vol. 12, no. 1, pp. A49–A57, Jan. 2020.
- [34] K. Guo, Y. Lu, H. Gao, and R. Cao, “Artificial intelligence-based semantic Internet of Things in a user-centric smart city,” *Sensors*, vol. 18, no. 5, p. 1341, May 2018.
- [35] D. Rafique and L. Velasco, “Machine learning for network automation: Overview, architecture, and applications [invited tutorial],” *J. Opt. Commun. Netw.*, vol. 10, no. 10, pp. D126–D143, Oct. 2018.
- [36] C. L. Stergiou, K. E. Psannis, and B. B. Gupta, “IoT-based big data secure management in the fog over a 6G wireless network,” *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5164–5171, Apr. 2021, doi: [10.1109/IOT.2020.3033131](https://doi.org/10.1109/IOT.2020.3033131).
- [37] D. Wang, B. Song, D. Chen, and X. Du, “Intelligent cognitive radio in 5G: AI-based hierarchical cognitive cellular networks,” *IEEE Wireless Commun.*, vol. 26, no. 3, pp. 54–61, Jun. 2019.
- [38] C. E. Shannon, “A mathematical theory of communication,” *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, 2001.

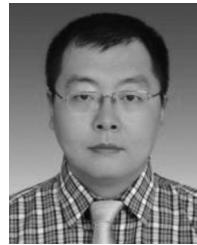
- [39] "The Mobile Broadband Standard." [Online]. Available: https://www.3gpp.org/ftp/Information/presentations/presentations_2020/Poster_2020_MWC_v6_OPTIMIZED.pdf (Accessed: Jul. 26, 2021).
- [40] L. Zhang, A. Ijaz, P. Xiao, A. Quddus, and R. Tafazolli, "Subband filtered multi-carrier systems for multi-service wireless communications," *IEEE Wireless Commun.*, vol. 16, no. 3, pp. 1893–1907, Jan. 2017.
- [41] P. Raviteja, K. T. Phan, Y. Hong, and E. Viterbo, "Interference cancellation and iterative detection for orthogonal time frequency space modulation," *IEEE Wireless Commun.*, vol. 17, no. 10, pp. 6501–6515, Oct. 2018.
- [42] R. Hadani *et al.*, "Orthogonal time frequency space modulation," in *Proc. IEEE WCNC*, San Francisco, CA, USA, Mar. 2017, pp. 1–6.
- [43] Z. Ding *et al.*, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.
- [44] G. Gui, H. J. Huang, Y. Song, and H. Sari, "Deep learning for an effective nonorthogonal multiple access scheme," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8440–8450, Sep. 2018.
- [45] N. Zhang, J. Wang, G. Kang, and Y. Liu, "Uplink nonorthogonal multiple access in 5G systems," *IEEE Commun. Lett.*, vol. 20, no. 3, pp. 458–461, Mar. 2016.
- [46] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO: Uniformly great service for everyone," in *Proc. IEEE SPAWC*, Stockholm Sweden, Jun. 2015, pp. 201–205.
- [47] D. Wang, Z. Zhao, Y. Huang, H. Wei, X. Wang, and X. You, "Large-scale multi-user distributed antenna system for 5G wireless communications," in *Proc. IEEE VTC-Spring*, Glasgow, U.K., May 2015, pp. 1–5.
- [48] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.
- [49] J. Zhang, S. Chen, Y. Lin, J. Zheng, B. Ai, and L. Hanzo, "Cell-free massive MIMO: A new next-generation paradigm," *IEEE Access*, vol. 7, pp. 99878–99888, 2019.
- [50] S. Buzzi and C. D'Andrea, "Cell-free massive MIMO: User-centric approach," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 706–709, Dec. 2017.
- [51] J. Zhang, Y. Wei, E. Björnson, Y. Han, and S. Jin, "Performance analysis and power control of cell-free massive MIMO systems with hardware impairments," *IEEE Access*, vol. 6, pp. 55302–55314, 2018.
- [52] Y. Ji, J. Zhang, Y. Xiao, and Z. Liu, "5G flexible optical transport networks with large-capacity, low-latency and high-efficiency," *China. Commun.*, vol. 16, no. 5, pp. 19–32, May 2019.
- [53] S. Zhou, X. Liu, F. Effenberger, and J. Chao, "Mobile-PON: A high-efficiency low-latency mobile fronthaul based on functional split and TDM-PON with a unified scheduler," in *Proc. OFC*, Los Angeles CA, USA, Mar. 2017, pp. 1–3.
- [54] M. Waqar, A. Kim, and J. J. Yoon, "A performance analysis of 5G fronthaul networks for long-distance communications," in *Proc. IEEE WD*, Manchester, U.K., Apr. 2019, pp. 1–4.
- [55] A. M. Mikaeil, W. Hu, T. Ye, and S. B. Hussain, "Performance evaluation of XG-PON based mobile front-haul transport in cloud-RAN architecture," *J. Opt. Commun. Netw.*, vol. 9, no. 11, pp. 984–994, Nov. 2017.
- [56] T. Nishitani, J. Mizuguchi, and H. Mukai, "Experimental study of type B protection for a TWDM-PON system," *J. Opt. Commun. Netw.*, vol. 7, no. 3, pp. A414–A420, Mar. 2015.
- [57] I. N. Cano *et al.*, "Experimental demonstration of a statistical OFDM-PON with multiband ONUs and elastic bandwidth allocation," *J. Opt. Commun. Netw.*, vol. 7, no. 1, pp. A73–A79, Jan. 2015.
- [58] W. Xie, N.-T. Mao, and K. Rundberget, "Cost comparisons of backhaul transport technologies for 5G fixed wireless access," in *Proc. IEEE 5GWF*, Silicon Valley, CA, USA, Jul. 2018, pp. 159–163.
- [59] L. Song and D. Hatzinakos, "Broadcasting energy efficiency limits in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 7, pp. 2502–2511, Jul. 2008.
- [60] L. Song and D. Hatzinakos, "A cross-layer architecture of wireless sensor networks for target tracking," *IEEE/ACM Trans. Netw.*, vol. 15, no. 1, pp. 145–158, Feb. 2007.
- [61] L. Song and D. Hatzinakos, "Cognitive networking of large scale wireless systems," *Int. J. Commun. Netw. Distrib. Syst.*, vol. 2, no. 4, pp. 452–475, Jun. 2009.
- [62] L. Song, P. Spachos, and D. Hatzinakos, "Large scale cognitive wireless networks: Architecture and application," in *Self-Organization and Green Applications in Cognitive Radio Networks*. Hershey, PA, USA: IGI Global, Jan. 2013, pp. 26–52.
- [63] P. Spachos, C. Periklis, and D. Hatzinakos, "Cognitive networking with opportunistic routing in wireless sensor networks," in *Proc. IEEE ICC*, Budapest, Hungary, Nov. 2013, pp. 2433–2437.
- [64] P. Spachos and D. Hatzinakos, "Scalable dynamic routing protocol for cognitive radio sensor networks," *IEEE Sensors J.*, vol. 14, no. 7, pp. 2257–2266, Jul. 2014.
- [65] F. Amirjavid, P. Spachos, L. Song, and K. N. Plataniotis, "Network coding in Internet of Things," in *Proc. IEEE CAMAD.*, Toronto, ON, Canada, Dec. 2016, pp. 140–145.
- [66] F. Amirjavid, P. Spachos, and K. N. Plataniotis, "3-D object localization in smart homes: A distributed sensor and video mining approach," *IEEE Syst. J.*, vol. 12, no. 2, pp. 1307–1316, Jun. 2018.
- [67] P. Spachos, L. Song, and K. N. Plataniotis, "Wireless noise prevention for mobile agents in smart home," in *Proc. IEEE ICC*, Paris, France, Jul. 2017, pp. 1–6.
- [68] P. T. Tin, D. T. Hung, T. N. Nguyen, T. T. Duy, and M. Voznak, "Secrecy performance enhancement for underlay cognitive radio networks employing cooperative multi-hop transmission with and without presence of hardware impairments," *Entropy*, vol. 21, no. 2, pp. 1–16, Jan. 2019.
- [69] D. Goel, V. S. Krishna, and M. Bhatnagar, "Selection relaying in decode-and-forward multi-hop cognitive radio systems using energy detection," *IET Commun.*, vol. 10, no. 7, pp. 753–760, May 2016.
- [70] D. Jiang, X. Ying, Y. Han, and Z. Lv, "Collaborative multi-hop routing in cognitive wireless networks," *Wireless Pers. Commun.*, vol. 86, no. 2, pp. 901–923, Sep. 2015.
- [71] A. Sheikholeslami, M. Ghaderi, D. Towsley, B. A. Bash, S. Guha, and D. Goecke, "Multi-hop routing in covert wireless networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3656–3669, Jun. 2018.
- [72] A. A. El-Banna, B. M. ElHalawany, A. B. Zaky, J. Huang, and K. Wu, "Machine learning-based multi-layer multi-hop transmission scheme for dense networks," *IEEE Commun. Lett.*, vol. 23, no. 12, pp. 2238–2242, Dec. 2019.
- [73] A. Y. Majid, M. Jansen, G. O. Delgado, K. S. Yildirim, and P. Pawełzak, "Multi-hop backscatter tag-to-tag networks," in *Proc. IEEE INFOCOM*, Paris, France, Apr. 2019, pp. 721–729.
- [74] B. Ying and A. Nayak, "A power-efficient and social-aware relay selection method for multi-hop D2D communications," *IEEE Commun. Lett.*, vol. 22, no. 7, pp. 1450–1453, Jul. 2018.
- [75] U. Challita and W. Saad, "Network formation in the sky: Unmanned aerial vehicles for multi-hop wireless backhauling," in *Proc. IEEE Global Commun. Conf.*, Singapore, Dec. 2017, pp. 1–6.
- [76] Z. M. Fadlullah *et al.*, "Multi-hop wireless transmission in multi-band WLAN systems: Proposal and future perspective," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 108–113, Feb. 2019.
- [77] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 388–404, Mar. 2000.
- [78] M. Grossglauser and D. N. C. Tse, "Mobility increases the capacity of ad hoc wireless networks," *IEEE/ACM Trans. Netw.*, vol. 10, no. 4, pp. 477–486, Aug. 2002.
- [79] N. Ophir, C. Mineo, D. Mountain, and K. Bergman, "Silicon photonic microring links for high-bandwidth-density, low-power chip I/O," *IEEE Micro*, vol. 33, no. 1, pp. 54–67, Jan./Feb. 2013.
- [80] S. Kumar *et al.*, "Scale MLPerf-0.6 models on Google TPU-v3 pods," Sep. 2019, *arXiv:1909.09756*.
- [81] S. Zhang *et al.*, "Cambricon-X: An accelerator for sparse neural networks," in *Proc. IEEE/ACM Micro*, Taipei, Taiwan, Oct. 2016, pp. 1–12.
- [82] F. Akopyan *et al.*, "TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 34, no. 10, pp. 1537–1557, Oct. 2015.
- [83] K. Liu and B. Wang, *Wireless AI: Wireless Sensing, Positioning, IoT, and Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2019.
- [84] C. Liu, B. Tian, and M. Ding, "Nanocomposite materials for nanoelectronic-based Internet of Things sensors and energy device signaling," in *Emerging 2D Materials and Devices for the Internet of Things*. Amsterdam, The Netherlands: Elsevier, Jun. 2020, pp. 243–290.
- [85] S. Sarkar, S. I. Alam, V. Ravindran, A. Banerjee, A. Debnath, and A. Ghosh, "Spintronics based systems for AI nano robots," in *Proc. GUCON*, Greater Noida, India, Sep. 2018, pp. 873–877.
- [86] A. P. James, "A hybrid memristor-CMOS chip for AI," *Nature Electron.*, vol. 2, no. 7, pp. 268–269, Jul. 2019.
- [87] W.-H. Chen *et al.*, "CMOS-integrated memristive non-volatile computing-in-memory for AI edge processors," *Nat. Electron.*, vol. 2, no. 9, pp. 420–428, Aug. 2019.

- [88] W. Zhang *et al.*, “Neuro-inspired computing chips,” *Nat. Electron.*, vol. 3, no. 7, pp. 371–382, Jul. 2020.
- [89] B. Murmann and B. Hoefflinger, “The new era of nano-chips: Green and intelligent,” in *NANO-CHIPS 2030*. Cham, Switzerland: Springer, 2020, pp. 1–7.
- [90] Y. Zhao, Y. Li, X. Zhang, G. Geng, W. Zhang, and Y. Sun, “A survey of networking applications applying the software defined networking concept based on machine learning,” *IEEE Access*, vol. 7, pp. 95397–95417, 2019.
- [91] C. Liu, A. Malboubi, and C.-N. Chuah, “OpenMeasure: Adaptive flow measurement & inference with online learning in SDN,” in *Proc. IEEE INFOCOM Workshop*, San Francisco, CA, USA, Apr. 2016, pp. 47–52.
- [92] A. Prajapati, A. Sakadasariya, and J. Patel, “Software defined Network: Future of networking,” in *Proc. IICISC*, Coimbatore, India, Jan. 2018, pp. 1351–1354.
- [93] “SDN Architecture Overview.” [Online]. Available: <https://opennetworking.org/wp-content/uploads/2013/02/SDN-architecture-overview-1.0.pdf> (Accessed: Oct. 17, 2021).
- [94] X. Huang, S. Cheng, K. Cao, P. Cong, T. Wei, and S. Hu, “A survey of deployment solutions and optimization strategies for hybrid SDN networks,” *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1483–1507, 2nd Quart., 2019.
- [95] R. Amin, M. Reisslein, and N. Shah, “Hybrid SDN networks: A survey of existing approaches,” *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3259–3306, 4th Quart., 2018.
- [96] Y. Li and M. Chen, “Software-defined network function virtualization: A survey,” *IEEE Access*, vol. 3, pp. 2542–2553, 2015.
- [97] “Network functions virtualisation,” ETSI, Sophia Antipolis, France, White Paper. Accessed: Jul. 6, 2021. [Online]. Available: https://portal.etsi.org/nfv/nfv_white_paper.pdf
- [98] A. J. Gonzalez, G. Nencioni, A. Kamisiński, B. E. Helvik, and P. E. Heegaard, “Dependability of the NFV orchestrator: State of the art and research challenges,” *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3307–3329, 4th Quart., 2018.
- [99] P. P. Ray and N. Kumar, “SDN/NFV architectures for edge-cloud oriented IoT: A systematic review,” *Comput. Commun.*, vol. 169, pp. 129–153, Mar. 2021.
- [100] C.-M. Huang, M.-S. Chiang, D.-T. Dao, W.-L. Su, S. Xu, and H. Zhou, “V2V data offloading for cellular network based on the software defined network (SDN) inside mobile edge computing (MEC) architecture,” *IEEE Access*, vol. 6, pp. 17741–17755, 2018.
- [101] R. Chaudhary, G. S. Aujla, N. Kumar, and J. J. P. C. Rodrigues, “Optimized big data management across multi-cloud data centers: Software-defined-network-based analysis,” *IEEE Commun. Mag.*, vol. 56, no. 2, pp. 118–126, Feb. 2018.
- [102] Q. Zhang, F. Liu, and C. Zeng, “Adaptive interference-aware VNF placement for service-customized 5G network slices,” in *Proc. IEEE INFOCOM*, Paris, France, Apr./May 2019, pp. 2449–2457.
- [103] Y. Li, L. T. X. Phan, and B. T. Loo, “Network functions virtualization with soft real-time guarantees,” in *Proc. IEEE INFOCOM*, Chengdu, China, Oct. 2016, pp. 1–9.
- [104] L. Li, K. Ota, and M. Dong, “DeepNFV: A lightweight framework for intelligent edge network functions virtualization,” *IEEE Netw.*, vol. 33, no. 1, pp. 136–141, Jan./Feb. 2019.
- [105] C.-T. Yang, S.-T. Chen, J.-C. Liu, Y.-Y. Yang, K. Mitra, and R. Ranjan, “Implementation of real-time network traffic monitoring service with network functions virtualization,” *Future Gener. Comput. Syst.*, vol. 93, pp. 687–701, Apr. 2019.
- [106] A. A. Barakabite, A. Ahmad, R. Mijumbi, and A. Hines, “5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges,” *Comput. Netw.*, vol. 167, Feb. 2020, Art. no. 106984.
- [107] P. Rost *et al.*, “Network slicing to enable scalability and flexibility in 5G mobile networks,” *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 72–79, May 2017.
- [108] R. A. Addad, T. Taleb, H. Flinck, M. Bagaa, and D. Dutra, “Network slice mobility in next generation mobile systems: Challenges and potential solutions,” *IEEE Netw.*, vol. 34, no. 1, pp. 84–93, Jan./Feb. 2020.
- [109] Y.-I. Choi and N. Park, “Slice architecture for 5G core network,” in *Proc. IEEE ICUFN*, Milan, Italy, Jul. 2017, pp. 571–575.
- [110] J. Backman, S. Yrjölä, K. Valtanen, and O. Mämmelä, “Blockchain network slice broker in 5G: Slice leasing in factory of the future use case,” in *Proc. Internet Things Bus. Models Users Netw.*, Copenhagen, Denmark, Nov. 2017, pp. 1–8.
- [111] D. Bega, M. Gramaglia, A. Garcia-Saavedra, M. Fiore, A. Banchs, and X. Costa-Perez, “Network slicing meets artificial intelligence: An AI-based framework for slice management,” *IEEE Commun. Mag.*, vol. 58, no. 6, pp. 32–38, Jun. 2020.
- [112] V. Sciancalepore, X. Costa-Perez, and A. Banchs, “RL-NSB: Reinforcement learning-based 5G network slice broker,” *IEEE/ACM Trans. Netw.*, vol. 27, no. 4, pp. 1543–1557, Aug. 2019.
- [113] X. Xu, H. Zhang, X. Dai, Y. Hou, X. Tao, and P. Zhang, “SDN based next generation mobile network with service slicing and trials,” *China Commun.*, vol. 11, no. 2, pp. 65–77, Feb. 2014.
- [114] H. Zhang, F. Fang, J. Cheng, K. Long, W. Wang, and V. C. M. Leung, “Energy-efficient resource allocation in NOMA heterogeneous networks,” *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 48–53, Apr. 2018.
- [115] X. Zhou, X. Wen, L. Wang, Z. Lu, and W. Guan, “SMDP-based resource allocation for slice requests with long-term reward maximization,” in *Proc. IEEE/CIC ICCC Workshops*, Changchun, China, Aug. 2019, pp. 36–41.
- [116] S. O. Oladejo and O. E. Falowo, “Latency-aware dynamic resource allocation scheme for multi-tier 5G Network: A network slicing-multitenancy scenario,” *IEEE Access*, vol. 8, pp. 74834–74852, 2020.
- [117] H. Halabian, “Distributed resource allocation optimization in 5G virtualized networks,” *IEEE J. Sel. Areas Commun.*, vol. 37, no. 3, pp. 627–642, Mar. 2019.
- [118] F. Xu, F. Yang, C. Zhao, and S. Wu, “Deep reinforcement learning based joint edge resource management in maritime network,” *China Commun.*, vol. 17, no. 5, pp. 211–222, May 2020.
- [119] T. Yang, Y. Guo, Y. Zhou, and S. Wei, “Joint communication and control for small underactuated USV based on mobile computing technology,” *IEEE Access*, vol. 7, pp. 160610–160622, 2019.
- [120] S. Shakottai, T. S. Rappaport, and P. C. Karlsson, “Cross-layer design for wireless networks,” *IEEE Commun. Mag.*, vol. 41, no. 10, pp. 74–80, Oct. 2003.
- [121] F. Tang *et al.*, “On removing routing protocol from future wireless networks: A real-time deep learning approach for intelligent traffic control,” *IEEE Wireless Commun.*, vol. 25, no. 1, pp. 154–160, Feb. 2018.
- [122] K. Rusek, J. Suárez-Varela, A. Metres, P. Barlet-Ros, and A. Cabellos-Aparicio, “Unveiling the potential of graph neural networks for network modeling and optimization in SDN,” in *Proc. Assoc. Comput. Mach.*, New York, NY, USA, Apr. 2019, pp. 140–151.
- [123] Q. Mao, F. Hu, and Q. Hao, “Deep learning for intelligent wireless networks: A comprehensive survey,” *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2595–2621, 4th Quart., 2018.
- [124] M. Pham, D. B. Hoang, and Z. Chaczko, “Congestion-aware and energy-aware virtual network embedding,” *IEEE/ACM Trans. Netw.*, vol. 28, no. 1, pp. 210–223, Feb. 2020.
- [125] C. Stergiou, K. E. Psannis, A. P. Plagieras, Y. Ishibashi, and B.-G. Kim, “Algorithms for efficient digital media transmission over IoT and cloud networking,” *J. Multimedia Inf. Syst.*, vol. 5, no. 1, pp. 27–34, Mar. 2018, doi: [10.9717/JMIS.2018.5.1.27](https://doi.org/10.9717/JMIS.2018.5.1.27).
- [126] K. E. Psannis, C. Stergiou, and B. B. Gupta, “Advanced media-based smart big data on intelligent cloud systems,” *IEEE Trans. Sustain. Comput.*, vol. 4, no. 1, pp. 77–87, Jan.–Mar. 2019.
- [127] A. Aghdai, M. I.-C. Wang, Y. Xu, C. H.-P. Wenz, and H. J. Chao, “In-network congestion-aware load balancing at transport layer,” in *Proc. IEEE NFV-SDN*, Dallas, TX, USA, Nov. 2019, pp. 1–6.
- [128] Y. Xu, W. Xu, Z. Wang, J. Lin, and S. Cui, “Load balancing for ultra-dense networks: A deep reinforcement learning-based approach,” *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9399–9412, Dec. 2019.
- [129] M. Li *et al.*, “Distributed machine learning load balancing strategy in cloud computing services,” *Wireless Netw.*, vol. 26, no. 8, pp. 5517–5533, Nov. 2020.
- [130] V. D. Chakravarthy and B. Amutha, “A novel software-defined networking approach for load balancing in data center networks,” *Int. J. Commun. Syst.*, vol. 35, no. 2, p. e4213, 2022, doi: [10.1002/dac.4213](https://doi.org/10.1002/dac.4213).
- [131] B. Niu, W. You, H. Tang, and X. Wang, “5G network slice security trust degree calculation model,” in *Proc. ICCC*, Chengdu, China, Dec. 2017, pp. 1150–1157.
- [132] D. Dwivedhika and T. Tachibana, “Virtual network embedding based on security level with VNF placement,” *Security Commun. Netw.*, vol. 2019, Feb. 2019, Art. no. 5640134, doi: [10.1155/2019/5640134](https://doi.org/10.1155/2019/5640134).
- [133] I. D. Alvarenga, G. A. F. Rebello, and O. C. M. B. Duarte, “Securing configuration management and migration of virtual network functions using blockchain,” in *Proc. IEEE/IFIP NOMS*, Taipei, Taiwan, Jul. 2018, pp. 1–9.

- [134] G. A. F. Rebello, I. D. Alvarenga, I. J. Sanz, and O. C. M. B. Duarte, “BSec-NFVO: A blockchain-based security for network function virtualization orchestration,” in *Proc. IEEE ICC*, Shanghai, China, Jul. 2019, pp. 1–6.
- [135] D. Silver *et al.*, “Mastering the game of Go with deep neural networks and tree search,” *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.
- [136] J. Dean *et al.*, “Large scale distributed deep networks,” in *Proc. NIPS*, Jan. 2012, pp. 1223–1231.
- [137] Y. Sheng and L. Zhou, “Distributed secure quantum machine learning,” *Sci. Bull.*, vol. 62, no. 14, pp. 1025–1029, Jul. 2017.
- [138] Y. Long, I. Chakraborty, G. Srinivasan, and K. Roy, “Complexity-aware adaptive training and inference for edge-cloud distributed AI systems,” in *Proc. IEEE 41st Int. Conf. Distrib. Comput. Syst. (ICDCS)*, 2021, pp. 573–583.
- [139] W. Y. B. Lin *et al.*, “Federated learning in mobile edge networks: A comprehensive survey,” *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 2031–2063, 3rd Quart., 2020.
- [140] Q. Meng, W. Chen, Y. Wang, Z. Ma, and T. Y. Liu, “Convergence analysis of distributed stochastic gradient descent with shuffling,” 2017, *arXiv:1709.10432*.
- [141] H. Zhu and Y. Jin, “Multi-objective evolutionary federated learning,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 4, pp. 1310–1322, Apr. 2020.
- [142] J. Dean and S. Ghemawat, “MapReduce: Simplified data processing on large clusters,” *Commun. ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008. [Online]. Available: <https://doi.org/10.1145/1327452.1327492>
- [143] E. P. Xing *et al.*, “Petuum: A new platform for distributed machine learning on big data,” *IEEE Trans. Big Data*, vol. 1, no. 2, pp. 49–67, Jun. 2015.
- [144] M. Li *et al.*, “Scaling distributed machine learning with the parameter server,” in *Proc. OSDI*, Broomfield, CO, USA, 2014, pp. 583–598.
- [145] J. K. Bradley, A. Kyrola, D. Bickson, and C. Guestrin, “Parallel coordinate descent for L_1 -regularized loss minimization,” in *Proc. ICML*, Bellevue, WA, USA, Jun. 2011, pp. 321–328.
- [146] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2010.
- [147] S. Zhang, A. Choromanska, and Y. Lecun, “Deep learning with elastic averaging SGD,” Dec. 2014, *arXiv:1412.6651*.
- [148] B. Recht, C. Re, S. J. Wright, and F. Niu, “HOGWILD!: A lock-free approach to parallelizing stochastic gradient descent,” in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran, 2011, pp. 693–701.
- [149] X. Pan *et al.*, “CYCLADES: Conflict-free asynchronous machine learning,” in *Proc. NIPS*, Barcelona, Spain, Dec. 2016 pp. 2576–2584.
- [150] Q. Ho *et al.*, “More effective distributed ML via a stale synchronous parallel parameter server,” in *Advanced in Neural Information Processing Systems*. Red Hook, NY, USA: Curran, 2013, pp. 1223–1231.
- [151] J. Konen, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, “Federated learning: Strategies for improving communication efficiency,” Oct. 2018, *arXiv:1610.05492*.
- [152] Z. Xu, Z. Yang, J. Xiong, J. Yang, and X. Chen, “Helios: Heterogeneity-aware federated learning with dynamically balanced collaboration,” Dec. 2019, *arXiv:1912.01684*.
- [153] C. Dinh *et al.*, “Federated learning over wireless networks: Convergence analysis and resource allocation,” 2019, *arXiv:1910.13067*.
- [154] T. Li, M. Sanjabi, A. Beirami, and V. Smith, “Fair resource allocation in federated learning,” 2020, *arXiv:1905.10497*.
- [155] G. Li, L. Liu, X. Wang, X. Dong, P. Zhao, and X. Feng, “Auto-tuning neural network quantization framework for collaborative inference between the cloud and edge,” in *Proc. ICANN*, Rhodes, Greece, Sep. 2018, pp. 402–411.
- [156] Y. Huang *et al.*, “Task scheduling with optimized transmission time in collaborative cloud-edge learning,” in *Proc. ICCCN*, Hangzhou, China, Aug. 2018, pp. 1–9.
- [157] X. Ran, H. Chen, X. Zhu, Z. Liu, and J. Chen, “DeepDecision: A mobile deep learning framework for edge video analytics,” in *Proc. IEEE INFOCOM*, Honolulu, HI, USA, 2018, pp. 1421–1429.
- [158] H.-J. Jeong, H.-J. Lee, C. H. Shin, and S.-M. Moon, “IONN: Incremental offloading of neural network computations from mobile devices to edge servers,” in *Proc. ACM Symp. Cloud Comput.*, Carlsbad, CA, USA, Oct. 2018, pp. 401–411.
- [159] M. Pipattanasomporn, H. Feroze, and S. Rahman, “Multi-agent systems in a distributed smart grid: Design and implementation,” in *Proc. IEEE/PES*, Seattle, WA, USA, Apr. 2009, pp. 1–8.
- [160] S. Gronauer and K. Diepold, “Multi-agent deep reinforcement learning: A survey,” *Artif. Intell. Rev.*, vol. 55, no. 2, pp. 895–943, 2022.
- [161] M. Raghu, A. Irpan, J. Andreas, R. Kleinberg, Q. V. Le, and J. M. Kleinberg, “Can deep reinforcement learning solve Erdos-Selfridge-Spencer games?” in *Proc. ICML PMLR*, Stockholm, Sweden, Jul. 2018, pp. 4238–4246.
- [162] J. K. Gupta, M. Egorov, and M. J. Kochenderfer, “Cooperative multi-agent control using deep reinforcement learning,” in *Proc. AAMAS*, vol. 10642. São Paulo, Brazil, May 2017, pp. 66–83.
- [163] J. Jiang and Z. Lu, “Learning attentional communication for multi-agent cooperation,” in *Proc. NIPS*, Montreal, QC, Canada, Dec. 2018, pp. 7265–7275.
- [164] H. Pham, M. Y. Guan, B. Zoph, Q. V. Le, and J. Dean, “Efficient neural architecture search via parameter sharing,” Feb. 2018, *arXiv:1802.03268*.
- [165] S. V. Albrecht and P. Stone, “Autonomous agents modelling other agents: A comprehensive survey and open problems,” *Artif. Intell.*, vol. 258, pp. 66–95, May 2018.
- [166] D. Mishra and E. Natalizio, “A survey on cellular-connected UAVs: Design challenges, enabling 5G/B5G innovations, and experimental advancements,” *Comput. Netw.*, vol. 182, Dec. 2020, Art. no. 107451.
- [167] R. Zhong, Y. Liu, X. Mu, Y. Chen, and L. Song, “AI empowered RIS-assisted NOMA networks: Deep learning or reinforcement learning?” *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 182–196, Jan. 2022.
- [168] “Autonomous Network Hyperloops.” [Online]. Available: <https://www.tmf.org/autonomous-network-hyperloops/> (Accessed: Jul. 22, 2021).
- [169] Y. Hao, Y. Miao, L. Hu, M. S. Hossain, G. Muhammad, and S. U. Amin, “Smart-edge-CoCaCo: AI-enabled smart edge with joint computation, caching, and communication in heterogeneous IoT,” *IEEE Netw.*, vol. 33, no. 2, pp. 58–64, Mar./Apr. 2019.
- [170] A. Roy, N. Saxena, B. J. R. Sahu, and S. Singh, “BiSON: A bio-inspired self-organizing network for dynamic auto-configuration in 5G wireless,” *Wireless Commun. Mobile Comput.*, vol. 2018, pp. 1–13, Nov. 2018.
- [171] Y. Ren, T. Phung-Duc, J.-C. Chen, and Z.-W. Yu, “Dynamic auto scaling algorithm (DASA) for 5G mobile networks,” in *Proc. GLOBECOM*, Washington, DC, USA, 2016, pp. 1–6.
- [172] M. G. Pérez *et al.*, “Dynamic reconfiguration in 5G mobile networks to proactively detect and mitigate botnets,” *IEEE Internet Comput.*, vol. 21, no. 5, pp. 28–36, Sep. 2017.
- [173] D. Bega, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, “DeepCog: Cognitive network management in sliced 5G networks with deep learning,” in *Proc. IEEE INFOCOM*, Paris, France, Apr. 2019, pp. 280–288.
- [174] T. T. Maksymyuk, J. Gazda, M. Ružička, E. Slapak, G. Bugar, and L. Han, “Deep learning based mobile network management for 5G and beyond,” in *Proc. TCSET*, Lviv-Slavskie, Ukraine, Feb. 2020, pp. 890–893.
- [175] H. Lee, S. H. Lee, and T. Q. S. Quek, “Deep learning for distributed optimization: Applications to wireless resource management,” *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2251–2266, Oct. 2019.
- [176] A. M. Zarca *et al.*, “Security management architecture for NFV/SDN-aware IoT systems,” *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8005–8020, Oct. 2019.
- [177] S. Wong, N. Sastry, O. Holland, V. Friderikos, M. Dohler, and H. Aghvami, “Virtualized authentication, authorization and accounting (V-AAA) in 5G networks,” in *Proc. IEEE CSCN*, Helsinki, Finland, Sep. 2017, pp. 175–180.
- [178] Z. Pan *et al.*, “Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts,” *Sci. China Inf. Sci.*, vol. 64, Nov. 2020, Art. no. 110301, doi: [10.1007/s11432-020-2955-6](https://doi.org/10.1007/s11432-020-2955-6).
- [179] A. Dorri, S. S. Kanhere, and R. Jurdak, “Towards an optimized BlockChain for IoT,” in *Proc. IEEE/ACM IoTDI*, Pittsburgh, PA, USA, Apr. 2017, pp. 173–178.
- [180] Y. Zhang, R. H. Deng, E. Bertino, and D. Zheng, “Robust and universal seamless handover authentication in 5G HetNets,” *IEEE Trans. Dependable Secure Comput.*, vol. 18, no. 2, pp. 858–874, Mar./Apr. 2021, doi: [10.1109/TDSC.2019.2927664](https://doi.org/10.1109/TDSC.2019.2927664).
- [181] A. Fitwi, Y. Chen, and S. Zhu, “A lightweight blockchain-based privacy protection for smart surveillance at the edge,” in *Proc. IEEE Int. Conf. Blockchain (Blockchain)*, Atlanta, GA, USA, Jul. 2019, pp. 552–555.

- [182] S. Kiyomoto, A. Basu, M. S. Rahman, and S. Ruj, "On blockchain-based authorization architecture for beyond-5G mobile services," in *Proc. ICITST*, Cambridge, U.K., Dec. 2017, pp. 136–141.
- [183] D. W. Kravitz and J. Cooper, "Securing user identity and transactions symbiotically: IoT meets blockchain," in *Proc. GIoTS*, Geneva, Switzerland, Jun. 2017, pp. 1–6.
- [184] S. H. Hashemi, F. Faghri, P. Rausch, and R. H. Campbell, "World of empowered IoT users," in *Proc. IoTDI*, Berlin, Germany, Apr. 2016, pp. 13–24.
- [185] M. A. Rashid and H. H. Pajooh, "A security framework for IoT authentication and authorization based on blockchain technology," in *Proc. IEEE TrustCom/BigDataSE*, Rotorua, New Zealand, Aug. 2019, pp. 264–271.
- [186] A. Pouraghily and T. Wolf, "A lightweight payment verification protocol for blockchain transactions on IoT devices," in *Proc. ICNC*, Honolulu, HI, USA, Feb. 2019, pp. 617–623.
- [187] M. F. Franco, E. J. Scheid, L. Z. Granville, and B. Stiller, "BRAIN: Blockchain-based reverse auction for infrastructure supply in virtual network functions-as-a -service," in *Proc. IFIP Netw.*, Warsaw, Poland, May 2019, pp. 1–9.
- [188] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *Proc. AAAI*, Quebec City, QC, Canada, Jul. 2014, pp. 1112–1119.
- [189] X. Chen, S. Jia, and Y. Xiang, "A review: Knowledge reasoning over knowledge graph," *Expert Syst. Appl.*, vol. 141, Mar. 2020, Art. no. 112948.
- [190] "Digital Twin." [Online]. Available: https://en.wikipedia.org/wiki/Digital_twin (Accessed: Dec. 21, 2021).
- [191] A. E. Saddik, "Digital twins: The convergence of multimedia technologies," *IEEE MultiMedia*, vol. 25, no. 2, pp. 87–92, Apr.–Jun. 2018.
- [192] D. Jones, C. Snider, A. Nassehi, J. Yon, and B. Hicks, "Characterising the digital twin: A systematic literature review," *CIRP J. Manuf. Sci. Technol.*, vol. 29, pp. 36–52, May 2020.
- [193] M. B. Chhetri, S. Krishnaswamy, and S. W. Loke, "Smart virtual counterparts for learning communities," in *Proc. WISE*, Brisbane, QLD, Australia, Nov. 2004, pp. 125–134.
- [194] G. Bondani and G. Bacchiega, "Creating an embedded digital twin: Monitor, understand and predict device health failure," presented at the Inn4mech Mechatronics Ind. 4.0 Conf. Ingegneria Ricerca Sistemi, Nov. 2019.
- [195] S. W. Loke *et al.*, "The La Trobe E-Sanctuary: Building a cross-reality wildlife sanctuary," in *Proc. IE*, Prague, Czech Republic, Aug. 2015, pp. 168–171.
- [196] "Azure Digital Twins." [Online]. Available: <https://azure.microsoft.com/en-us/services/digital-twins/> (Accessed: Jul. 19, 2021).
- [197] F. Tao, H. Zhang, A. Liu, and A. Y. C. Nee, "Digital twin in industry: State-of-the-art," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2405–2415, Apr. 2019.
- [198] M. Trancossi, M. Cannistraro, and J. Pascoa, "Can constructal law and exergy analysis produce a robust design method that couples with industry 4.0 paradigms? the case of a container house," *Math. Model. Eng. Problems*, vol. 5, no. 4, pp. 303–312, May 2018.
- [199] "MRO 4.0: The Next Big Step for Processes—Using Digital Twins and Location." [Online]. Available: <https://www.aircraftit.com/articles/improving-business-efficiency-in-an-mro-environment/> (Accessed: Jul. 23, 2021).
- [200] "Immersive Technology." [Online]. Available: https://en.wikipedia.org/wiki/Immersive_technology (Accessed: Jul. 24, 2021).
- [201] J. Dorrier, "The Next Great Computer Interface is Emerging—But It Doesn't Have a Name Yet." [Online]. Available: <https://singularityhub.com/2017/05/21/the-next-great-computer-interface-is-emerging-but-it-doesnt-have-a-name-yet> (Accessed: Jun. 13, 2021).
- [202] M. J. Maas and J. M. Hughes, "Virtual, augmented and mixed reality in K-12 education: A review of the literature," *Technol. Pedagogy Educ.*, vol. 29, no. 2, pp. 231–249, 2020.
- [203] "Vive Shipment Updates—VIVE Blog." [Online]. Available: <https://blog.vive.com/us/2016/04/07/vive-shipment-updates/> (Accessed: Aug. 11, 2021).
- [204] H. Sumra, "HTC Vive Pro: Everything You Need to Know About the 3K Headset." [Online]. Available: <https://www.wareable.com/vr/htc-vive-pro-release-date-price-specs-995> (Accessed: Oct. 24, 2021).
- [205] J. Martindale, "Vive-Like Sensor Spotted in New Sony Patent Could Make Its Way to PlayStation VR." [Online]. Available: <https://www.digitaltrends.com/virtual-reality/sony-psvr-patent-sensor/> (Accessed: Oct. 25, 2021).
- [206] "From the Lab to the Living Room: The Story Behind Facebook's Oculus Insight Technology and a New Era of Consumer VR." [Online]. Available: <https://tech.fb.com/the-story-behind-oculus-insight-technology/> (Accessed: Dec. 16, 2021).
- [207] "Virtual Reality." [Online]. Available: https://en.wikipedia.org/wiki/Virtual_reality (Accessed: Jul. 20, 2021).
- [208] J. Martinez, D. Griffiths, V. Biscione, O. Georgiou, and T. Carter, "Touchless haptic feedback for supernatural VR experiences," in *Proc. IEEE VR*, Reutlingen, Germany, Mar. 2018, pp. 629–630.
- [209] "Augmented Reality." [Online]. Available: https://en.wikipedia.org/wiki/Augmented_reality (Accessed: Oct. 9, 2021).
- [210] "The History of Augmented Reality." Huffington Post. [Online]. Available: https://www.huffpost.com/entry/the-history-of-augmented-_b_9955048 (Accessed: Jul. 14, 2021).
- [211] N. M. B. Tuanquin, S. Hoermann, C. J. Petersen, and R. W. Lindeman, "The effects of olfactory stimulation and active participation on food cravings in virtual reality," in *Proc. IEEE Conf. Virtual Reality 3D User Interfaces*, Reutlingen, Germany, Mar. 2018, pp. 709–710.
- [212] H.-K. Wu, S. W.-Y. Lee, H.-Y. Chang, and J.-C. Liang, "Current status, opportunities and challenges of augmented reality in education," *Comput. Educ.*, vol. 62, pp. 41–49, Mar. 2013.
- [213] J. Steuer, "Defining virtual reality: Dimensions determining telepresence," *J. Commun.*, vol. 42, no. 4, pp. 73–93, Dec. 1992.
- [214] "Immersing Consumers in 'Immersive Experiences.'" [Online]. Available: www.mashable.com (Accessed: Jul. 17, 2021).
- [215] N. Mohammadi and J. E. Taylor, "Smart city digital twins," in *Proc. IEEE SSCI*, Honolulu, HI, USA, Nov. 2017, pp. 1–5.
- [216] K. Ahir, K. Govani, R. Gajera, and M. Shah, "Application on virtual reality for enhanced education learning, military training and sports," *Augmented Human Res.*, vol. 5, no. 1, pp. 1–9, 2020.
- [217] D. Beck *et al.*, "Immersive learning research network," in *Proc. iLRN*, Coimbra, Portugal, 2019, p. 247.
- [218] N. Ancombe, "A different view: In Europe, virtual and immersive technologies are starting to change the landscape of medical simulation," *IEEE Pulse*, vol. 6, no. 4, pp. 14–19, Jul./Aug. 2015.
- [219] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," Aug. 2015, *arXiv:1508.01991*.
- [220] O. Etzioni *et al.*, "Unsupervised named-entity extraction from the Web: An experimental study," *Artif. Intell.*, vol. 165, no. 1, pp. 91–134, Jun. 2005.
- [221] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open information extraction from the Web," *Commun. ACM*, vol. 51, no. 12, pp. 68–74, Dec. 2008.
- [222] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proc. ACL*, Singapore, Aug. 2009, pp. 1003–1011.
- [223] S. Zheng *et al.*, "Joint entity and relation extraction based on a hybrid neural network," *Neurocomputing*, vol. 257, pp. 59–66, Sep. 2017.
- [224] S. Zheng, F. Wang, H. Bao, Y. Hao, P. Zhou, and B. Xu, "Joint extraction of entities and relations based on a novel tagging scheme," Jun. 2017, *arXiv:1706.05075*.
- [225] X. Zhao, Y. Jia, A. Li, R. Jiang, and Y. Song, "Multi-source knowledge fusion: A survey," *World Wide Web*, vol. 23, no. 4, pp. 2567–2592, 2020.
- [226] Z. Xu, G. L. Frankwick, and E. Ramirez, "Effects of big data analytics and traditional marketing analytics on new product success: A knowledge fusion perspective," *J. Bus. Res.*, vol. 69, no. 5, pp. 1562–1566, 2016.
- [227] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. NIPS*, Dec. 2013, pp. 2787–2795.
- [228] Z. Tian *et al.*, "Real-time lateral movement detection based on evidence reasoning network for edge computing environment," *IEEE Trans. Ind. Informat.*, vol. 15, no. 7, pp. 4285–4294, Jul. 2019.
- [229] S. Wang, J. Wan, D. Li, and C. Liu, "Knowledge reasoning with semantic data for real-time data processing in smart factory," *Sensors*, vol. 18, no. 2, p. 471, Feb. 2018.
- [230] X. Jiang, S. Du, Z. Qin, Y. Sun, and J. Yu, "KBGN: Knowledge-bridge graph network for adaptive vision-text reasoning in visual dialogue," Aug. 2020, *arXiv:2008.04858*.
- [231] H. Chen and X. Luo, "An automatic literature knowledge graph and reasoning network modeling framework based on ontology and natural language processing," *Adv. Eng. Informat.*, vol. 42, Oct 2019, Art. no. 100959.

- [232] X. Chen, J. Xie, J. Ji, and Z. Sui, "Toward open knowledge enabling for human-robot interaction," *J. Human-Robot Interact.*, vol. 1, no. 2, pp. 100–117, Jan. 2013.
- [233] W. Liu *et al.*, "Enabling language representation with knowledge graph," Sep. 2019, *arXiv:1909.07606*.
- [234] M. Farsi, A. Daneshkhah, A. Hosseiniyan-Far, and H. Jahankhani, *Digital Twin Technologies and Smart Cities*. Cham, Switzerland: Springer, 2020.
- [235] N. Shaukat *et al.*, "A survey on electric vehicle transportation within smart grid system," *Renew. Sustain. Energy Rev.*, vol. 81, no. 1, pp. 1329–1349, Jan. 2018.
- [236] A. Sumalee and H. W. Ho, "Smarter and more connected: Future intelligent transportation system," *IATSS Res.*, vol. 42, no. 2, pp. 67–71, Jul. 2018.
- [237] H. Magsi, A. H. Sodhro, F. A. Chachar, S. A. K. Abro, G. H. Sodhro, and S. Pirbhulal, "Evolution of 5G in Internet of medical things," in *Proc. IEEE iCoMET*, Sukkur, Pakistan, Feb. 2018, pp. 1–7.
- [238] Y. Yin, Y. Zeng, X. Chen, and Y. Fan, "The Internet of Things in healthcare: An overview," *J. Ind. Inf. Integr.*, vol. 1, pp. 3–13, Mar. 2016.
- [239] S. Gul *et al.*, "A survey on role of Internet of Things in education," *Int. J. Comput. Sci. Netw. Security*, vol. 17, no. 5, pp. 159–165, May 2017.
- [240] Y. Lu, "Industry 4.0: A survey on technologies, applications and open research issues," *J. Ind. Inf. Integr.*, vol. 6, pp. 1–10, Jun. 2017.
- [241] J. Wan *et al.*, "Software-defined industrial Internet of Things in the context of industry 4.0," *IEEE Sensors J.*, vol. 16, no. 20, pp. 7373–7380, Oct. 2016.
- [242] M. S. Mekala and P. Viswanathan, "A survey: Smart agriculture IoT with cloud computing," in *Proc. IEEE ICMDCS*, Vellore, India, Aug. 2017, pp. 1–7.
- [243] S. A. Oke, "A literature review on artificial intelligence," *Int. J. Inf. Manag. Sci.*, vol. 19, no. 4, pp. 535–570, Dec. 2018.
- [244] H. Lund, P. A. Ostergaard, D. Connolly, and B. V. Mathiesen, "Smart energy and smart energy systems," *Energy*, vol. 137, pp. 556–565, Oct. 2017.
- [245] S. E. Bibri, "The eco-city and its core environmental dimension of sustainability: Green energy technologies and their integration with data-driven smart solutions," *Energy Informat.*, vol. 3, no. 1, pp. 1–26, 2020.
- [246] N. Kosmyna, J. T. Lindgren, and A. Lécuyer, "Attending to visual stimuli versus performing visual imagery as a control strategy for EEG-based brain-computer interfaces," *Science*, vol. 9, no. 8, pp. 2–3, Sep. 2018.
- [247] B.-K. Min, S. Dähne, M.-H. Ahn, Y.-K. Noh, and K.-R. Müller, "Decoding of top-down cognitive processing for SSVEP-controlled BMI," *Nature*, vol. 9, no. 6, p. 2, Nov. 2016.
- [248] M. I. Sereno *et al.*, "Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging," *Science*, vol. 268, no. 5212, pp. 889–893, May 1995.
- [249] C. W. Tyler, L. T. Likova, C.-C. Chen, L. L. Kontsevich, M. M. Schira, and A. R. Wade, "Extended concepts of occipital retinotopy," *Current Med. Imag. Rev.*, vol. 1, no. 3, pp. 319–329, 2005.
- [250] B. A. Wandell, A. A. Brewer, and R. F. Dougherty, "Visual field map clusters in human cortex," *Philos. Trans. Roy. Soc. London B, Biol. Sci.*, vol. 360, no. 1456, pp. 693–707, Apr. 2005.
- [251] J. V. Haxby *et al.*, "Dissociation of object and spatial visual processing pathways in human extra striate cortex," in *Proc. Nat. Acad. Sci. USA*, vol. 88, no. 5, pp. 1621–1625, Mar. 1991.
- [252] M. Pihlajamäki, H. Taniela, M. Könönen, T. Hänninen, H. J. Aronen, and H. Soininen, "Distinct and overlapping fMRI activation networks for processing of novel identities and locations of objects," *Eur. J. Neurosci.*, vol. 22, no. 8, pp. 2095–2105, Oct. 2005.
- [253] T. Xie, D. Zhang, Z. Wu, L. Chen, and X. Zhu, "Classifying multiple types of hand motions using electrocorticography during intraoperative awake craniotomy and seizure monitoring processes—Case studies," *Front. Neurosci.*, vol. 9, p. 353, Oct. 2015.
- [254] L. R. Hochberg *et al.*, "Neuronal ensemble control of prosthetic devices by a human with tetraplegia," *Nature*, vol. 442, no. 7099, pp. 164–172, Jul. 2016.
- [255] "About Neuro-Link." [Online]. Available: <https://neurolink.company/cms/about-neuro-link/> (Accessed: Dec. 20, 2021).
- [256] L. Jiang, A. Stocco, D. M. Losey, J. A. Abernethy, C. S. Prat, and R. P. N. Rao, "BrainNet: A multi-person brain-to-brain interface for direct collaboration between brains," *Nat. Sci. Rep.*, vol. 9, pp. 1–11, Apr. 2019.
- [257] "What Are Digital Twins?—IBM Developer." [Online]. Available: <https://developer.ibm.com/technologies/iot/articles/what-are-digital-twins> (Accessed: Jul. 20, 2021).
- [258] "VR, AR, MR: Which Reality Technology to Choose for Your Business in 2020." [Online]. Available: <https://rubygarage.org/blog/difference-between-ar-vr-mr> (Accessed: Dec. 13, 2021).
- [259] P. R. Srivastava and S. Saurav, "Networking agent for overlay L2 routing and overlay to underlay external networks L3 routing using OpenFlow and open vSwitch," in *Proc. APNOMS*, Busan, South Korea, Aug. 2015, pp. 291–296.
- [260] N. Cheng *et al.*, "A comprehensive simulation platform for space-air-ground integrated network," *IEEE Wireless Commun.*, vol. 27, no. 1, pp. 178–185, Feb. 2020.
- [261] C. Stergiou, K. E. Psannis, B.-G. Kim, and B. Gupta, "Secure integration of IoT and cloud computing," *Future Gener. Comput. Syst.*, vol. 78, pp. 964–975, Jan. 2018.
- [262] C. Stergiou, K. E. Psannis, B. B. Gupta, and Y. Ishibashi, "Security, privacy & efficiency of sustainable cloud computing for big data & IoT," *Sustain. Comput. Inform. Syst.*, vol. 19, pp. 174–184, Sep. 2018.



Liang Song (Senior Member, IEEE) received the B.E. degree from Shanghai Jiao Tong University, Shanghai, China, in 1999, the M.S. degree from Fudan University, Shanghai, in 2002, and the Ph.D. degree from the University of Toronto, Toronto, ON, Canada, in 2006.

He is a Professor with Fudan University, along with numerous distinguished or adjunct appointments, such as the Vice Chairman of the Institute on Networking Systems of AI, the Professor with the University of Toronto, and the Chairman of Shanghai 5G-VR Alliance. He is currently the Director of Fudan Institute on Networking Systems of AI, Fudan University, with more than 30 faculties. His research interests converge communication networks and AI systems, where he has published more than 100 referred papers, and invented over 100 patents.

Dr. Song is a Fellow of the Canadian Academy of Engineering and the Chinese National Distinguished Experts.



Xing Hu received the Ph.D. degree in control science and engineering from Shanghai Jiao Tong University, Shanghai, China, in 2016.

He is an Associate Professor with the School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai. His current research interests include image processing, computer vision, machine learning, and biomedical signal processing.



Guanhua Zhang received the Ph.D. degree in computer science from Shanghai Jiao Tong University, Shanghai, China, in 2005.

He is currently a Research Fellow with Fudan Institute on Networking Systems of AI, Fudan University, Shanghai, and the Vice Dean of Shanghai East-Bund Institute on Networking System of AI, Shanghai. His current research interests include networking, embedded system, and AI.



Petros Spachos (Senior Member, IEEE) received the Diploma degree in electronic and computer engineering from the Technical University of Crete, Chania, Greece, in 2008, and the M.A.Sc. and Ph.D. degrees in electrical and computer engineering from the University of Toronto, Toronto, ON, Canada, in 2010 and 2014, respectively.

He is currently an Associate Professor with the School of Engineering, University of Guelph, Guelph, ON, Canada. He is also a Registered Professional Engineer in Ontario. His research interests include experimental wireless networking and mobile computing with a focus on wireless sensor networks, smart cities, and Internet of Things.



Konstantinos (Kostas) N. Plataniotis (Fellow, IEEE) received the B.Eng. degree in computer engineering from the University of Patras, Patras, Greece, in 1988, and the M.S. and Ph.D. degrees in electrical engineering from the Florida Institute of Technology Melbourne, FL, USA, in 1991 and 1994, respectively.

He is a Professor and the Bell Canada Chair in Multimedia with the University of Toronto, Toronto, ON, Canada. He is a Registered Professional Engineer in Ontario. His research interests are in the areas of machine learning and signal processing, and their applications in imaging systems, communications systems, and knowledge media design systems.

Dr. Plataniotis is the General Co-Chair of the 2027 IEEE International Conference on Acoustics, Speech and Signal Processing. He is a Fellow of the Engineering Institute of Canada and the Canadian Academy of Engineering/L' Academie Canadienne Du Genie.



Hequan Wu received the B.Eng. degree from the Department of Wireline Communications, Wuhan Post and Telecommunications Institute, Wuhan, China, in 1964.

He is a renowned expert in optical fiber transmission network and broadband information network. His early contributions include the research and development of digital communication equipment and optical fiber transmission systems. Over past three decades, he has successively undertaken or organized the Chinese National 863 Program in Communication High Technology, the 973 Key Fundamental Research Program, the Next-generation Internet Pilot Project, and the New-generation Broadband Wireless Mobile Communication Network Program. He has also been responsible for multiple major Chinese national consulting projects in science and technology.

Dr. Wu is an Academician of the Chinese Academy of Engineering. He had been the Vice President and the Chief Engineer of the China Academy of Telecommunications Technology, and the Vice President of the Chinese Academy of Engineering. He is currently the Director of the Chinese National Standardization Expert Advisory Committee, the Chinese IPv6 Scale-Deployment Expert Committee, and the Advisory Committee of the Chinese Internet Society.