Reinforcement Learning Based Physical Cross-Layer Security And Privacy in 6G

Xiaozhen Lu, *Member, IEEE,* Liang Xiao, *Senior Member, IEEE,* Pengmin Li, Xiangyang Ji, *Member, IEEE,* Chenren Xu, *Senior Member, IEEE,* Shui Yu, *Senior Member, IEEE,* Weihua Zhuang, *Fellow, IEEE*

***Abstract*—Sixth-generation (6G) cellular systems will have an inherent vulnerability to physical (PHY)-layer attacks and privacy leakage, due to the large-scale heterogeneous networks with booming time-sensitive applications. Important wireless techniques including non-orthogonal multiple access, mobile edge computing, millimeter-wave, massive multiple-input and multiple-output, visible light communication, terahertz, and intelligent reflecting surface can improve the spectrum efficiency and quality-of-service but will raise challenges for the 6G PHY and cross-layer security and privacy protection. Existing optimization based PHY and cross-layer security and privacy protection schemes such as the convex optimization method have to rely on accurate attack patterns and strategies and thus suffer from performance degradation in 6G systems that have shorter communication latency, more devices and higher spectrum efficiency than 5G. Reinforcement learning (RL) algorithms help wireless devices optimize their security policies to enhance the security performance in dynamic networks against smart attacks without depending on the attack model. Therefore, this article provides a comprehensive survey on the RL based 6G PHY cross-layer security and privacy protection. In this article, we investigate the potential attacks in 6G systems and discuss the PHY cross-layer security solutions. A brief overview of reinforcement learning algorithms is provided. Afterward, we review the RL based PHY-layer security and privacy protection and discuss how to apply RL algorithms in 6G security scenarios, especially focusing on the game with jammers, eavesdroppers, spoofers and inference attackers. The RL based security solutions for unmanned aerial vehicles (UAVs) and cross-layer scenarios are also reviewed. The future research directions are identified and the corresponding RL based potential solutions are discussed for 6G.**

***Index Terms*—6G, PHY-layer security, privacy, reinforcement learning, secure communications, UAVs, cross-layer security.**

I. Introduction

Sixth-generation (6G) cellular systems will increase the network connectivity and coverage to support flourishing real-time Internet of Everything (IoE) applications such as telemedicine, tactile Internet, autonomous driving, intelligence disaster prediction, and surreal virtual reality (VR) compared with 5G [1]. Thus, the 6G security performance will be degraded by the physical (PHY)-layer attacks (e.g., jamming, spoofing, and eavesdropping) and the network attacks such as denial-of-service (DoS) attacks, man-in-the-middle attacks, and selfish attacks [2]–[4]. With more strict communication latency (e.g., 10 $\sim$ 100 μs in [5]), 10 times more Internet of Things (IoT) devices and 5 $\sim$ 10 times higher spectrum efficiency than 5G, 6G will have more challenges in the PHY-layer security and privacy protection [6], [7]. In particular, the communication techniques as 6G candidates, including non-orthogonal multiple access (NOMA) [8], mobile edge computing (MEC) [9], visible light communication (VLC) [10], terahertz (THz) [11], intelligent reflecting surface (IRS) [12], millimeter-wave (mmWave) [13], and massive multiple-input and multiple-output (MIMO), have to address the security threats, such as jamming attacks in VLC and THz systems [14], [15], and selfish attacks in MEC [16].

By exploiting the physical properties of the communication channels or the unforgeable features of the device hardware components, PHY-layer security solutions such as anti-jamming [17], secure communications [18] and PHY-layer authentication [19] incorporate with the higher-layer security techniques such as cryptography and authentication in [20] to enhance the security performance against the PHY-layer attacks. Compared with the higher-layer security techniques, PHY-layer security makes use of the time-varying and random nature of wireless channels instead of the encryption keys to enhance the security performance and thus reduce the communication and computational overhead in cellular systems [21]. For example, a lightweight anti-jamming massive MIMO system in [17] applies convex optimization to optimize the base station (BS) transmit power under the Gaussian channel and jamming model to improve the achievable rate against reactive jamming attacks in wireless networks. The PHY-layer spoofing detection in [19] applies the maximum likelihood estimation to discriminate the radio transmitters in the Rayleigh fading channel model but the detection accuracy degrades in dynamic wireless networks.

As another important technique for 6G, the cross-layer security designs the solutions at both the PHY-layer and the

This work was supported in part by the Natural Science Foundation of China under Grants U21A20444, 62202222, 61971366, and U22B2062, and the Natural Science Foundation of Jiangsu Province under Grant BK20220880. (*Corresponding author: Liang Xiao*)

X. Lu was with the Department of Information and Communication Engineering, Xiamen University during this work, and is now with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China. E-mail: luxiaozhen@nuaa.edu.cn.

L. Xiao and P. Li are with the Department of Information and Communication Engineering, Xiamen University, Xiamen 361005, China. E-mail: lxiao@xmu.edu.cn.

X. Ji is with the Department of Automation, Tsinghua University, Beijing 100084, China. Email: xyji@tsinghua.edu.cn.

C. Xu is with the Department of Computer Science and Technology, Peking University, Beijing 100871, China. Email: chenren@pku.edu.cn.

S. Yu is with the Faculty of Engineering and Information Technology, University of Technology Sydney, Australia. Email: Shui.Yu@uts.edu.au.

W. Zhuang is with Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L3G1, Canada. Email: wzhuang@uwaterloo.ca.

higher layers (such as the link layer, the network layer, and the application layer), to resist the cross-layer attacks including eavesdropping [22], distributed DoS (DDoS) attacks [23], and man-in-the-middle attacks [24]. For instance, the security-aware cross-layer scheme in [22] uses the dual decomposition method to formulate a secrecy energy-efficient maximization problem following the average packet dropping probability and total energy consumption constraints against the passive eavesdropper that can steal the data at the media access control (MAC) layer and the PHY-layer. However, most of the PHY-layer and the cross-layer security solutions rely on the accurate attack patterns or strategies, thus suffering from performance degradation with an unknown attack model, and increasing heterogeneity and complexity of 6G systems, under smart attackers that can change their attack strategies intelligently.

Reinforcement learning (RL) algorithms as a potential solution, such as Q-learning, policy hill climbing (PHC), Dyna-Q, post-decision state (PDS) and double-Q [25], enable wireless devices to quickly optimize their security policies based on observations of the environment, instead of relying on an accurate attack model as in the optimization-based security methods or the labeled training data as in supervised learning. Thus, they can improve the security and privacy protection performance of dynamic 6G systems against more intelligent PHY-layer and cross-layer attackers [26]. For example, the secure healthcare transmission in [27] that applies both Dyna-Q and PDS to optimize the offloading policy exploits the state that consists of the sensing data size, the battery level and the radio channel state to improve the privacy level without being aware of the eavesdropping strategies. However, these tabular RL based security schemes have slower learning efficiency under a large feasible action set with a large number of mobile users and access points (APs), and the required optimization time increases with the network scale, the number of antennas and the amount of the available communication resource.

Being successfully applied in the video and strategy board games such as Alpha-Go, deep RL including deep Q-network (DQN) [28], asynchronous advantage actor-critic (A3C) [29], deep deterministic policy gradient (DDPG) [30] and proximal policy optimization (PPO) [31] designs deep neural networks (DNNs), including convolutional neural networks (CNNs) and recurrent neural networks, to accelerate the learning speed of PHY-layer security in large-scale systems. For instance, the secure VLC system in [32] applies DDPG to choose the beamforming vector and significantly increases the secrecy rate against passive eavesdropping. Nevertheless, the deep RL based schemes depend on the accurate state and reward signals after performing each security policy and the sufficient computational resources to support deep learning, which are not always available in cellular systems [33].

There are instructive surveys on the RL based 5G and beyond techniques such as power control, computational offloading and edge caching in [34]–[36], the machine learning based 6G security techniques in [5], [6] and the 6G security and privacy in [1], [37]. Particularly, the 6G security survey in [6] focuses on unsupervised learning and deep learning instead of RL and addresses the poisoning attacks and reverse attacks instead of the PHY-layer attacks. Another 6G security survey

TABLE I
SUMMARY OF ABBREVIATIONS

Abbreviations	**Definitions**
6G	Sixth-generation
IoE	Internet of Everything
VR	Virtual reality
PHY	Physical
DoS	Denial-of-service
IoT	Internet of Things
NOMA	Non-orthogonal multiple access
MEC	Mobile edge computing
VLC	Visible light communication
THz	Terahertz
IRS	Intelligent reflecting surface
MmWave	Millimeter-wave
MIMO	Multiple-input and multiple-output
BS	Base station
DDoS	Distributed denial of service
MAC	Media access control
RL	Reinforcement learning
PHC	Policy hill climbing
PDS	Post-decision state
AP	Access point
DQN	Deep Q-network
A3C	Asynchronous advantage actor-critic
DDPG	Deep deterministic policy gradient
PPO	Proximal policy optimization
DNN	Deep neural network
CNN	Convolutional neural network
UAV	Unmanned aerial vehicle
RFID	Radio-frequency identification
QoS	Quality of service
MDP	Markov decision process
SARSA	State-action-reward-state-action
MCS	Mobile crowdsensing
NEC	Neural episodic control
DDQN	Double DQN
A2C	Advantage actor-critic
DND	Differentiable neural dictionary
DIAL	Differentiable inter-agent learning
MADDPG	Multi-agent deep deterministic policy gradient
BER	Bit error rate
CMDP	Constrained MDP
SINR	Signal-to-interference-plus-noise ratio
Conv.	Convolutional
FC	Fully connected
AN	Artificial noise
RSSI	Received signal strength indicator
DP	Differential privacy
PoW	Proof of work
PSNR	Peak signal-to-noise ratio
SDN	Software-defined networking
MAML	Model-agnostic meta-learning

in [1] provides a brief overview of the security and privacy at the PHY-layer, the connection layer and the service layer and discusses the impact of the artificial intelligence on 6G systems rather than the application of RL in the PHY-layer and cross-layer security and privacy protection scenarios.

In this article, we review important potential security and privacy leakage threats for 6G and discuss the RL algorithms that have been applied to enhance wireless security. We investigate RL based security and privacy protection schemes and show how to apply RL algorithms to optimize the 6G security policy, such as the security resource allocation and authentication parameters against jamming, eavesdropping, spoofing, inference attacks and selfish attacks. This article focuses on the RL based PHY-layer security for 6G systems with NOMA, MEC, VLC, THz, IRS, mmWave and massive MIMO, and points out the remaining challenges to be addressed. As a newly developed technique, the RL based security and privacy for unmanned aerial vehicles (UAVs) is reviewed, including the anti-jamming communications, secure communications and data privacy-aware communications. We also discuss the application of RL in PHY cross-layer security solutions for 6G and the corresponding challenges. The security performance degradation due to the partial observation, dangerous exploration, communication overhead, and high curse of dimensionality of the 6G security scenarios can be addressed by the promising solutions including meta-learning, transfer learning, federated learning, safe RL and multi-agent RL.

The remaining part of this survey is organized as follows. We summarize the 6G potential security threats in Section II and review the typical RL algorithms in Section III, followed by the RL based PHY-layer security and privacy protection solutions in Sections IV-VIII. We discuss the future research directions in Section IX and summarize this survey in Section X. The abbreviations of this article are summarized in Table I.

II. Security Threats in 6G Systems

Due to more restricted requirements (e.g., higher mobility, data rate, spectrum efficiency, network energy efficiency and area traffic capacity) and booming real-time IoE applications such as UAVs, 6G systems with communication techniques such as VLC, THz and IRS are more vulnerable to PHY-layer attacks including jamming, identity-based attacks, and eavesdropping, and the higher-layer attacks (e.g., DoS attacks, man-in-the-middle attacks, and selfish attacks) than the existing cellular systems [38]. Besides, the attackers that exist in 4G or 5G are more harmful to 6G systems due to the more open, distributed, and intelligent networks. As shown in Fig. 1, the attackers can intercept the legitimate signals in VLC, THz, and mmWave, send jamming or spoofing signals to block the transmission of NOMA, massive MIMO, MEC and IRS, and thus result in DoS attacks, and inject fake messages to cause severe privacy leakage.

A. Jamming

Jammers send replayed, fake or random signals to block the ongoing signals between the mobile users and BSs or APs with the goal of degrading their transmission quality, depleting the device energy, and launching further attacks such as DoS attacks and man-in-the-middle attacks [39]. Typical jamming includes proactive jamming and reactive jamming.

1) Proactive jamming: Jammers send jamming signals to degrade the message reception performance without observing the ongoing transmission status. Proactive jamming contains constant jamming, sweep jamming and random jamming.

- **Constant jamming:** A jammer uses a radio device such as a waveform generator to send jamming signals on a number of frequency channels at some specific time but suffers from jamming energy waste in the absence of the victim transmissions [40].
- **Random jamming:** The jamming signals are sent on a frequency band randomly chosen in random time duration. The random jamming pattern makes this type of jamming more energy-efficient compared with constant jamming.
- **Sweep jamming:** A jammer switches channels periodically and simultaneously and sends jamming signals in the next periodic instead of jamming immediately. This type of jamming results in a large number of retransmissions and thus exhausts the device energy [41].

2) Reactive jamming: A reactive jammer uses a radio receiver to observe the ongoing transmission states and sends jamming signals according to the spectrum sensing results to improve the jamming efficiency. For example, an attacker sends jamming signals on the channel that has the preamble or pilots of the legitimate packets. The pilot contamination as a special case can significantly reduce the achievable rate for massive MIMO systems [17].

- **Smart jamming:** By using smart radio devices such as the universal software radio peripheral, an attacker can observe the ongoing transmission and induce wireless devices to apply the specifical defense strategy, whose communication performance can be improved by RL such as Q-learning and DQN that optimizes the jamming frequency and power levels. For example, the RL based smart jammers use the observed channel status as the basis to optimize the jamming power to degrade the transmission performance of NOMA systems [42]. Another smart jammer as designed in [43] builds a deep learning classifier to predict the next transmission with the previously observed channel transmission information and then jams the system accordingly.

B. Identity-based Attacks

Attackers impersonate legitimate mobile users with their identities, abuse multiple user identities, and generate pseudonymous identities or manipulate fake identities to obtain illegal access to the mobile users and BSs/APs [15]. This type of attack such as spoofing and Sybil attacks can send spoofing signals such as wrong reports or spam to fool users, emulate a large number of legitimate users to prevent legal access and thus launch further attacks, e.g., replay attacks and DoS attacks.

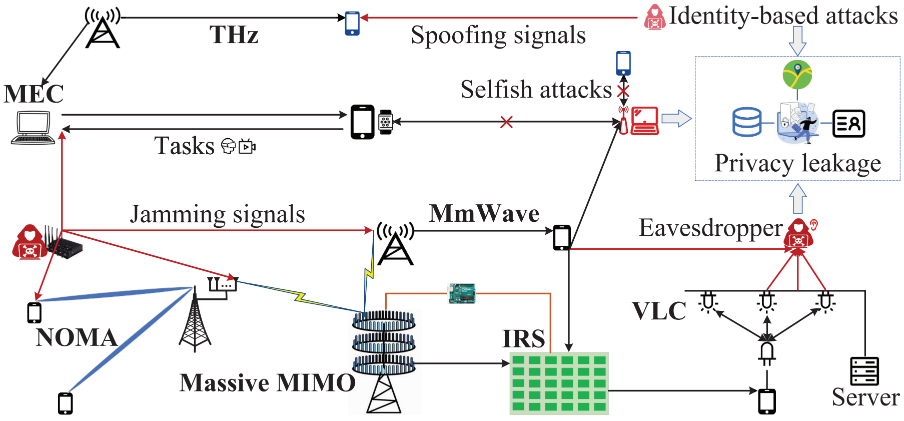


Fig. 1. Security threats and privacy model in 6G systems.

- **Spoofing attacks:** The attacker sends spoofing messages with the fake identities of the other wireless devices, such as the MAC addresses, radio-frequency identification (RFID) tags and IP addresses of the other mobile users, BSs and APs to obtain illegal advantages [44]. As illustrated in Fig. 1, a fake edge attacker impersonates the edge device to send fake computational results and the rogue AP attacker aims to access the data of the mobile users.
- **Sybil attacks:** This type of attacker impersonates a number of users to win majority votes and obtain other advantages, which will cause message collisions, wrong warnings and privacy leakage in 6G systems [45]. For example, the Sybil attacker with a large number of pseudonymous identities spreads spam and advertisements and even disseminates malware and fishing websites to other users for stealing user private information in IoT systems [46].

C. Eavesdropping

Eavesdroppers can intercept or steal data during the VLC, THz and mmWave transmission, including passive and active eavesdropping [15].

- **Passive eavesdropping:** The attacker aims to intercept the ongoing data without degrading the message reception performance of users and applies the traffic analysis to infer user privacy such as the communication pattern and the user location based on the number of the transmitted packets, the inter-packet duration and the traffic directionality [47].
- **Active eavesdropping:** The active eavesdropper not only receives the legal signals but also uses radio devices to send jamming signals to interrupt the ongoing message reception, raise the device transmit power and thus steal more data to infer user private information [48].

D. DoS Attacks

DoS attackers continuously send service requests or jamming signals to flood the servers and edge devices to prevent legal mobile devices from obtaining 6G network services [49]. In particular, a DDoS attacker impersonates a large number of network devices with their IP addresses, injects the forged service request messages continuously and generates malicious traffic with the mobile botnets to flood the unnecessary requests and interrupt the legal services [50]. As illustrated in Fig. 1, the attacker compromises a number of mobile devices and controls them to send multiple computing requests to the edge device and thus shut down the edge computing services.

E. Man-in-the-middle Attacks

The attacker eavesdrops on the transmission status in the communication channel, intercepts the ongoing transmitted messages between two users, modifies the intercepted messages and injects the manipulated messages into the 6G systems to deceive or even control radio devices [51], which aims to fool the receiver and thus obtain some user sensitive information to gain illegal profits [52]. As shown in Fig. 1, the attacker in mmWave systems intercepts the messages of the transmitter, replaces the intercepted messages with fake messages and then sends them to the receiver.

F. Selfish Attacks

Selfish users and edge devices sometimes refuse to help relay to save their limited bandwidth, energy, buffer resources and privacy, and manipulate the service records of the other devices in the reputation-based 6G systems for their own interests [53]. Potential impact includes the increased transmission latency and energy consumption, the degraded transmission quality and the service failure [54]. For example, a selfish edge device sends false computational results to the mobile devices or uses less computational resources than promised to save its computational resources [55].

G. Privacy Leakage

The 6G systems will be required to protect the privacy leakage from linking attacks, inference attacks, differential attacks and reconstruction attacks [56]. According to [57], user privacy consists of identity, data and location privacy.

- **Identity privacy:** The user private identity information such as the name, home address, phone number and private keys is easily exposed to attackers due to the increased number of mobile users [58]. For example, a man-in-the-middle attacker can falsify the request of a mobile user to obtain the user private key.

- **Data privacy:** The increasing data-intensive applications in 6G systems such as the smart metering and healthcare services make the user data such as the body sensing data expose to malicious attacks [2]. For instance, an inference attacker applies association rules and Bayesian reasoning to analyze the house status or economic status based on the intercepted meter consumption data.
- **Location privacy:** The user location can indicate the user behavior, preferences, personal habits and beliefs [59]. For example, an adversary can analyze the frequency and duration of the visit to the hospital to infer the type of illnesses of users, sell the health information for illegal profits and sometimes even cause crimes.

In summary, 6G systems need to address PHY-layer attacks such as jamming, spoofing and eavesdropping as well as higher-layer attacks such as man-in-the-middle attacks, selfish attacks, linking attacks, inference attacks, differential attacks and reconstruction attacks. With the rapid advancement in artificial intelligence, smart attackers can observe the defense states and improve the attack patterns accordingly, raising new challenges to security in 6G systems.

III. Overview and Tutorial of Reinforcement Learning

Machine learning techniques (such as supervised learning, unsupervised learning, and RL) will enable dynamic and heterogeneous 6G systems to improve the security and privacy protection performance [5], [16], [60]. In particular, supervised learning (such as support vector machine and logistic regression) has been applied in [5] to improve the intrusion detection accuracy, but the performance degrades in the dynamic systems without sufficient training data. Unsupervised learning such as K-means has been applied in the anomaly detection, as in [61], [62], to improve the detection accuracy without relying on any training data, but cannot always satisfy the quality of service (QoS) requirements of security applications. On the other hand, the RL based security solutions do not rely on the labeled dataset or prior knowledge of the attack and network model [25].

Mostly developed for video and strategy board games such as Alpha-Go, RL is promising to enhance wireless security for UAV networks, vehicular ad hoc networks, IoT systems and cellular systems [60]. In a finite Markov decision process (MDP), RL maintains the Q-table that outputs the long-term expected reward (i.e., Q-values), the mixed policy values that output the policy probability, and DNNs that output the Q-values, the state values or the advantage values, as shown in Fig. 2. More specifically, the learning agent (such as a smartphone or a wireless device) observes the environment that contains the wireless communication systems (e.g., mmWave, IRS, THz and VLC systems), the PHY-layer attackers such as jammers and eavesdroppers and the higher-layer attackers such as DoS attackers. The observation is used to formulate the state that represents the network status and security performance (such as the outage probability and the packet loss ratio) resulting from the previous actions and states. The learning agent uses the state as the basis to optimize its action (such as the security protocol and parameters) via trial-and-error in the learning process, and receives the reward such as the transmission quality or the privacy protection level from the environment after performing the chosen action.

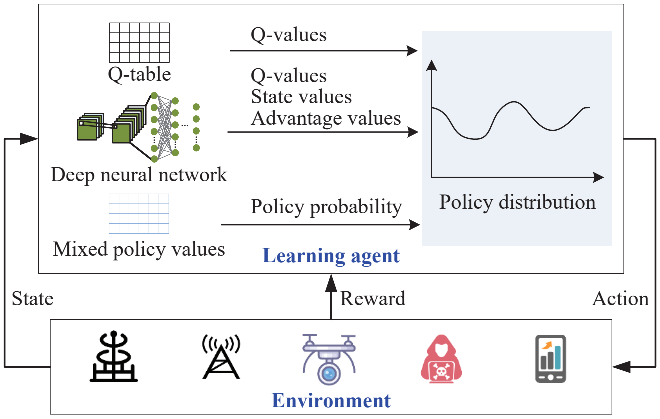


Fig. 2. Illustration of reinforcement learning in wireless communications.

Typical single-agent RL algorithms containing the tabular-based RL and deep RL have been proved to enable the agent to maximize the long-term expected reward without relying on the known network model of the dynamic environments and the attacker policy. More specifically, the tabular-based RL algorithms (such as Q-learning) use arrays or tables with one entry for each state-action pair to obtain approximations of value functions such as the Q-values and the mixed policy values of the learning agents in the small-scale systems with finite and discrete state space and action set. On the other hand, the deep RL includes the value-based RL that learns the value function (e.g., the Q-values) based on the temporal difference learning and the policy gradient RL that learns the policy distribution or mixed policies from the neural network function approximators [63].

A. *Tabular-based Reinforcement Learning*

Important tabular-based RL algorithms in wireless security applications include Q-learning, PHC, Dyna-Q, PDS, state-action-reward-state-action (SARSA)-Q, and double-Q. As illustrated in Fig. 3, the value functions such as the Q-values or the mixed policy values are the basis to choose the security action for the given state. These algorithms usually work well under low-dimensional discrete action-state space.

- **Q-learning:** As the first model-free RL algorithm that does not rely on the transition probability distribution associated with the corresponding MDP, Q-learning depends on a fixed policy distribution or the ϵ-greedy algorithm in the action selection from a small action set and updates the Q-values with an iterative Bellman equation every time slot [64]. For example, the mobile device with limited computational resources can apply

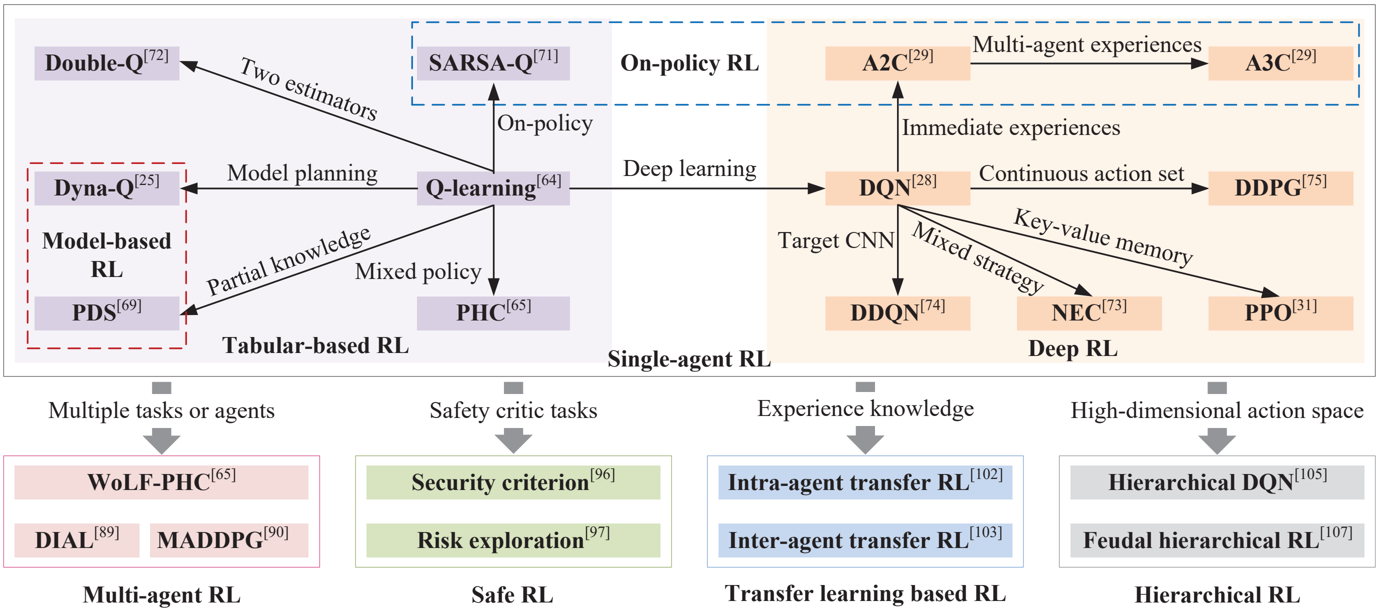


Fig. 3. Important reinforcement learning algorithms.

Q-learning to optimize the security policy in a discrete action set (such as the transmit power level in anti-jamming communications and the authentication mode in PHY-layer authentication systems) under the discrete state space. Without relying on any known model, Q-learning suffers from a slow learning speed under dynamic and complex environments, and the optimization policy that maximizes the Q-values can be easily estimated by the attacker that knows enough.

- **PHC:** As an extension of Q-learning, PHC combines the mixed policy values that represent the probability to select an action in the learning process with Q-values, and introduces randomness in the action selection to fool attackers for wireless security applications [65]. Each mixed policy value is initialized with the average of all the feasible actions and updated in each time slot by increasing the probability of the action that is expected to maximize the Q-values with a weighted value, and decreasing the other probabilities with the weighted value averaged over all the feasible actions. For instance, the UAV as an agent applies PHC to choose the relay policy with mixed policy and thus fool the smart jammer in the vehicular communication system [66].
- **Dyna-Q:** As a model-based RL, Dyna-Q designs a Dyna architecture to simulate virtual learning experiences for planning and provide more policy learning [25]. More specifically, the Dyna architecture simulates a number of state-action pairs from the action set and state space in every time slot and calculates the corresponding reward to obtain the virtual learning experiences. Both the virtual learning experiences and the real experiences are used to update the Q-values, yielding faster policy optimization than Q-learning. Dyna-Q has been used to improve the learning efficiency in the selection of the power allocation in the anti-jamming NOMA communication, the authentication mode and test threshold in the PHY-layer authentication [44], [67], the network coding policy in the secure communication [68], and the offloading policy in the privacy-aware communication [27].
- **PDS:** By exploiting the environment knowledge such as the channel state transition, PDS saves the unnecessary learning samples or interactions with the environment based on the known information such as the channel gain of the legitimate users and thus accelerates learning in complicated networks with large state spaces [69]. For instance, vehicles can apply PDS to optimize their sensing strategy based on the known channel variance in dynamic mobile crowdsensing (MCS) systems against selfish attacks [70].
- **SARSA-Q:** As an on-policy temporal difference algorithm, SARSA-Q selects the action similar to Q-learning while updating the Q-values with the current state-action pair rather than the maximum Q-values of the next state-action pair [71]. The wireless devices with insufficient energy and computational resources (e.g., the IoT devices) can use SARSA-Q to improve the privacy protection performance for 6G systems against inference attacks, in which the Q-values of the available privacy policies under the current state are used to formulate the policy distribution.
- **Double-Q:** Double-Q uses two estimators to randomly update one of the two Q-tables, in order to reduce the over-estimation of the Q-values and thus avoids achieving suboptimal policies [72]. For example, the mobile devices in PHY-layer authentication systems can quantize the test threshold from zero to 1 and choose the authentication mode and the test threshold with the chosen Q-table. However, it requires a larger storage space compared with Q-learning.

In summary, the tabular-based RL algorithms have been

TABLE II
COMPARISON OF RL ALGORITHMS

RL algorithms	Mixed policy value	Model-free	On-policy	Continuous action set	Value-based	Policy gradient
Q-learning [64]	×	✓	×	×	✓	×
PHC [65]	✓	✓	×	×	✓	×
Dyna-Q [25]	×	×	×	×	✓	×
PDS [69]	×	×	×	×	✓	×
SARSA-Q [71]	×	✓	✓	×	✓	×
Double-Q [72]	×	✓	×	×	✓	×
DQN [28]	×	✓	×	×	✓	×
NEC [73]	×	✓	×	×	✓	×
DDQN [74]	×	✓	×	×	✓	×
A2C [29]	✓	✓	✓	✓	×	✓
A3C [29]	✓	✓	✓	✓	×	✓
PPO [31]	✓	✓	✓	✓	×	✓
DDPG [75]	×	✓	×	✓	✓	✓

analyzed thoroughly and their efficacy in the MDPs with relatively low-dimensional action-state space has been verified in wireless security applications [42], [44], [55]. Based on the discrete action-state set, these algorithms have to address the performance degradation due to the state quantization errors, especially in a complicated dynamic 6G system under multiple attacks.

B. Deep Reinforcement Learning

By combining deep learning, deep RL applies DNNs such as CNNs to extract the state features and compress the high-dimensional state-action space for higher learning efficiency in heterogeneous 6G systems. As summarized in Table II, the value-based RL includes DQN, neural episodic control (NEC), double DQN (DDQN), and the policy gradient RL consists of advantage actor-critic (A2C), A3C, and PPO. In particular, DDPG belongs to both the value-based RL and the policy gradient RL. All these deep RL algorithms are important for wireless security applications.

- **DQN:** Based on a CNN to compress the state space and a target CNN to improve the policy selection stability under the highly correlated state, DQN applies the ϵ-greedy to evaluate the outputs of the former CNN, i.e., the Q-values of the feasible policies under the current state [28]. Different from Q-learning, this algorithm helps wireless devices with sufficient computational resources choose the security policy in a discrete action set without quantizing the state space and thus improves the security performance in a more complicated system.
- **NEC:** In this algorithm, an online CNN generates keys of the differentiable neural dictionary (DND) (i.e., a key-value memory module), instead of directly generating Q-values for all the available policies under the current state. The previous experiences in the DNDs are exploited to reduce the learning sample complexity compared with DQN [73]. As an example, NEC can be applied in the NOMA system with a large number of multi-antenna users to choose the transmission policy that consists of the user subchannel and the BS transmit power against active eavesdropping, in which the BS involves several DNDs with size equaling to the feasible transmission policies.
- **DDQN**: Compared with DQN, this algorithm designs a target network with the same network architecture as the online network to evaluate the target Q-values of the action that maximizes the long-term expected reward under the next state and thus reduces the probability to achieve the local optimal policies [74]. For example, DDQN can enable the MEC system in [76] to optimize the edge selection, offloading and caching policies, and the NOMA system with a large number of users and BS with multiple antennas in [77] to choose the number of spreading codes without relying on the known attack interval.
- **A2C:** As an on-policy RL algorithm, A2C consists of a critic network that estimates the value function of the state and an actor network to estimate the advantage function of the action and output the policy distribution [29]. Both the two neural networks are updated with the immediate experiences rather than the previous experiences. This algorithm uses the mixed policy and introduces randomness in the action selection to fool the adversary, e.g., in the mmWave anti-jamming system with a large number of mmWave propagation channels, the BS can apply A2C to optimize the beamforming policy to fool the jammer with omni-directional antennas and thus improve the communication performance.
- **A3C:** This algorithm uses a global network and multiple subnetworks to improve the policy optimization speed and thus reduces the sample and computational complexity [29]. A global network that samples a number of experiences from the subnetworks is used to update the weights asynchronously to increase the diversity of training data. Each subnetwork copies the weights from the global network to update their networks independently. For instance, the MEC server that has enough resources to

support deep learning can apply A3C to optimize the user offloading policy and transmit power, and the security strategy such as the block size and thus protect the user data privacy.

- **PPO:** As a mixed policy and on-policy algorithm that can deal with both the discrete and continuous action set [31], PPO applies Kullback-Leibler divergence to evaluate the chosen policies and uses the importance sampling technique to update the network weights with the previous experiences instead of randomly sampling from the memory pool to reduce the storage overhead compared with DQN. The mobile devices with limited storage resources in 6G cross-layer security applications such as the wireless area body networks and MEC systems can apply PPO to optimize the friendly jamming, the offloading policy and the cross-layer security policy such as the encryption key length of data based on the output multivariate Gaussian policy distribution against active eavesdropping [78], [79].
- **DDPG:** An online actor network is used to directly generate the action rather than the policy distribution to avoid the estimation errors, and an online critic network is used to update the weights of the online actor network to avoid tracking suboptimal policies [30], [75]. Different from PPO that outputs the policy distribution and updates the network parameters, this algorithm uses two target networks with the sampling experiences from the memory pool to provide more experiences to update the online networks, and thus avoids the instability exploration in the learning process. By dealing with the continuous action set such as the UAV yaw angle in [80], this algorithm reduces the action quantization errors of the value-based RL such as DQN and NEC. As another important application, the DDPG based PHY-layer security system directly outputs the anti-jamming or secure communication policies (such as the transmit power, encryption key size, and mobility strategy), thus improving the policy optimization accuracy of the PPO based approach in [79].

The deep RL algorithms enable the learning agent that has sufficient computational resources to support deep learning to improve the PHY-layer security performance in complicated systems. These algorithms have been applied to choose the transmit power and channel, the authentication mode and parameters, the encryption key parameters and the protection level [81]–[86].

C. Multi-agent Reinforcement Learning

Multi-agent RL enables multiple learning agents such as wireless devices to share their experiences with similar tasks by communication, teaching or imitation for faster learning and better performance in large-scale networks [87]. Multi-agent RL algorithms such as win or learn fast PHC (WoLF-PHC), differentiable inter-agent learning (DIAL) and multi-agent deep deterministic policy gradient (MADDPG) reduce the task failure for the agents newly entering the system due to the experience exchange with the existing agents.

- **WoLF-PHC:** By combining PHC with the win or learn fast principle, WoLF-PHC depends on the average policy distribution of all the agents in the learning process [65]. The learning agent updates the policy distribution by adding the probability of current action with the winning learning rate ranging from zero to one if the probability to choose the action is larger than the average probability over all the feasible actions, and updates it with the losing learning rate otherwise. Each learning agent chooses its action independently without communication with the other agents. For example, the UAV and the vehicle in the anti-jamming air-to-ground communication system such as [88] apply WoLF-PHC to choose their transmit power levels with mixed policies based on the historical transmission quality, which aims to induce the smart attacker to use the incorrect jamming policies that result in higher overhead.
- **DIAL:** Deep recurrent Q-networks are used to reduce the storage in a large-scale system and the agents share their policy gradient parameters and observations to avoid the backpropagation errors and thus reduce the unnecessary random policy exploration [89]. This algorithm enables each agent to use a Q-network to generate the Q-values and use a communication network to share the current observations or the previous chosen actions to other agents with feedback signals. For instance, the PHY-layer authentication system in [44] can apply DIAL to choose the authentication mode and test threshold from the action set based on the shared experiences of each mobile device. The reward function can be formulated as the sum of the authentication accuracy minus the latency of all the mobile devices.
- **MADDPG:** This algorithm shares the observations and actions among the agents rather than the network weights in the network update process [90]. Each agent uses its own observations as the input of the actor network that directly outputs the action, and uses the observations and the actions of the other agents as the input of the critic network that updates the weights of the actor network. As a potential application, MADDPG can help improve the secure communication performance in the selection of the friendly jamming power levels for the massive MIMO system with hundreds of multi-antenna users against eavesdropping.

The multi-agent RL provides additional experiences and distributed hyperparameter tuning to achieve collaborative intelligence for agents with sufficient computing and storage resources in large-scale systems such as device-to-device underlay cellular networks and MEC-assisted vehicular ad hoc networks. A number of multi-agent RL algorithms are applied to optimize the user scheduling, the UAV trajectory, the transmit power of users, and the spectrum, computing and caching resources [91]–[94].

D. Safe Reinforcement Learning

Safe RL uses a security criterion in the reward, including the worst-case criterion, the risk-sensitive criterion and the constrained criterion. By designing a risk modulation such as providing initial knowledge, teacher advice and risk-directed exploration in the policy selection process, this RL algorithm avoids choosing the risky actions that result in network security disaster (e.g., the communication failure [95] and the device damage [96]).

- **Security criterion in the reward:** The RL algorithm incorporates the security factor such as the outage probability in the anti-jamming communications and designs a punishment term such as the variance of the Q-values, the probability to an error state, and the security constraints in the reward for the dangerous policies that fail to finish the tasks. For example, the RL with convex constraints in [96] incorporates the constraints as a penalty signal into the reward function and uses policy gradient methods to optimize the policy under the security constraints such as rock avoidance in the Mars Rover game. Safe RL can also help optimize the cross-layer security policy without exploring vulnerable policies that cannot satisfy the security constraints such as the QoS requirements. For instance, the mobile device uses an outage probability constraint as the security criterion to design a punishment function in the selection of the encryption key size and transmit power [79].
- **Risk based exploration:** The security criterion, such as the risk level of each state-action pair in terms of the security performance metrics, is evaluated and used as the basis to formulate the security policy distribution [97]. For example, safe DQN in [95] designs a security criterion based on the outage probability or the bit error rate (BER) of each state-action pair to evaluate the long-term expected risk values (i.e., E-values) and formulates a Boltzmann policy distribution based on both the long-term expected risk values and Q-values.

Safe RL enables the mobile devices to choose their policy with the goal of maximizing the long-term expected reward in the constrained MDP (CMDP) with various security constraints such as the task computational latency requirements in MEC [98]. Nevertheless, the wireless devices have to incorporate the policies of other devices to design an appropriate security criterion in the learning exploration for a large-scale system with multiple various tasks.

E. Transfer Learning Based Reinforcement Learning

Transfer learning algorithms such as intra-agent transfer RL and inter-agent transfer RL [99] enable RL to exploit the experiences in similar scenarios to initialize or update the learning parameters for faster policy optimization [100]. For example, a progressive RL algorithm enables the mobile device to learn the security experiences of a set of tasks and abstract the knowledge to a higher-level representation to initialize the Q-values and the DNN weights [101].

- **Intra-agent transfer RL:** A learning agent uses the experiences in the same or similar tasks to initialize the learning parameters [102]. For example, a wireless device in [42], [66] initializes the long-term expected reward and the mix-strategy policy distribution based on the reward, action and state of the previous similar anti-jamming tasks, which improves the learning speed without sharing experiences with the other agents.
- **Inter-agent transfer RL:** A learning agent integrates its own experiences with the experiences shared by the other agents in similar tasks to accelerate learning for the large-scale network. For example, the transfer learning framework in [103] uses an abstract knowledge base for the agent that executes an underlying task to extract the previous actions of the similar tasks and uses an advisor to provide additional learning experiences for the agent.

Transfer RL algorithms help improve the learning efficiency in multi-task or multi-agent systems such as MEC and cell-free massive MIMO systems [55]. However, the intra-agent transfer RL fails to exploit the experiences of the other agents performing the same task in the environment. Besides, the exchange of experiences among agents sometimes causes a high communication overhead and even user data privacy leakage.

F. Hierarchical Reinforcement Learning

Hierarchical RL such as hierarchical DQN and feudal hierarchical RL applies a hierarchical control architecture to exploit the temporal abstraction, in order to compress the large action set and improve the exploration efficiency for the semi-MDPs [104].

- **Hierarchical DQN:** The multi-goal problem in DQN suffers from sparse and delayed reward signals over high-dimensional state spaces, which can be addressed by using the hierarchical action-value functions or Q-values. More specifically, the feasible actions are divided into two sub-actions: the first sub-action is chosen based on the first layer Q-values, and the second sub-action is selected according to the second layer Q-values and the chosen first sub-action [105]. For instance, the top-level of the PHY-layer authentication scheme in [106] inputs the state of the detection accuracy and the message priority, and outputs the top Q-values to choose the authentication mode. The bottom-level inputs the state and the chosen mode that is the authentication basis, and outputs the bottom Q-values to select the authentication parameter.
- **Feudal hierarchical RL:** By combining the hierarchical structure that divides the task into two levels (i.e., a top level and a lower level) with the feudal RL that deals with a multi-goal task scenario and consists of several managers and workers, this algorithm uses a manager in the top level and a worker in the lower level to optimize the policy, and applies a differentiable neural network with two levels of hierarchy to improve the back-propagation efficacy. The CNN is replaced with dilated long short-term memory in the top-level of the manager for higher learning efficiency in the long-term multi-goal tasks [107]. In this algorithm, the manager sets goals for the top level at a lower temporal resolution in a delayed state-space, and the worker follows the goal of the manager. Specifically, the manager calculates the state representation and uses an approximate transition policy

TABLE III
RL BASED ANTI-JAMMING COMMUNICATION SCHEMES

Learning agent	Policy	Reward	RL algorithms	Applications
BS	Power allocation	SINR User sum rate	Q-learning Dyna-Q [42]	NOMA
	Power allocation Antenna selection Subband selection	User sum rate Energy consumption	Hierarchical DQN [108]	
Mobile device	Edge selection Offloading rate Power allocation	BER Computational latency Energy consumption	Safe-Q DDPG [98]	MEC
BS	Power allocation IRS phase shifts	Energy consumption Outage probability Achivable data rate	WoLF-PHC [109]	IRS
UAV	UAV location IRS beamforming vector	Transmission rate Computational overhead	DDPG [110]	
BS	Beamforming vector Power allocation	BER User sum rate Outage probability	A2C [111]	MmWave
BS	Power allocation	SINR User sum rate Energy consumption	PHC [112]	Massive MIMO
User		Achievable rate Energy efficiency	DQN [17]	
		Sum rate Energy consumption	MADDPG [113]	

gradient to train its abstract goal in a lower resolution that is used to guide the policy selection of the worker. With the external observation, state and goal of the manager as the input, the worker chooses the action at a higher temporal resolution.

Hierarchical RL solves the course of dimensionality issues in typical RL algorithms and improves the policy optimization efficiency. Nevertheless, how to incorporate the policy priority in the hierarchical structure is critical for wireless devices whose security policies have different priorities.

In summary, the tabular-based RL such as Q-learning, Dyna-Q and PDS can improve the security performance for small-scale networks with discrete action set and state space, deep RL further improves the learning efficiency in more complicated and dynamic networks with high-dimensional state space, and multi-agent RL enables each wireless devices to fast choose their security policies in the large-scale networks. On the other hand, safe RL will help reduce dangerous exploration for 6G security and further avoid communication failure and device damage. The learning speed of the typical deep RL can be accelerated by transfer learning based RL and hierarchical RL. The basis for choosing the RL algorithms in a 6G security and privacy protection scenario includes the network scale, attack capacity, dimensions of the state space, resources of learning agents, QoS requirements, etc.

IV. RL BASED ANTI-JAMMING COMMUNICATIONS

RL based radio resource allocation such as the transmission power allocation helps improve the anti-jamming communication performance for 6G systems with NOMA, MEC, IRS, mmWave, and massive MIMO, as summarized in Table III.

A. NOMA

NOMA systems support BSs to transmit messages to all the users in the same frequency subband, which are vulnerable to jamming. Convex optimization and Lagrangian optimization enable the BS or user to optimize the power allocation and the channel selection against jamming, based on the accurate jammer channel state, locations and patterns [108], which are rarely known by practical NOMA systems. The RL based anti-jamming NOMA communication is optimized based on the state, which can be the channel states, the data rate and the transmission quality such as the signal-to-interference-plus-noise ratio (SINR), instead of the jamming model. The reward depends on the BER of the received messages, the throughput, the packet loss rate, the outage probability and the energy consumption.

For example, the Q-learning based NOMA power allocation in [42] optimizes the power allocation for each user with multiple antennas, based on the state containing the SINR of the received signal instead of the accurate jamming model and strategies as in the optimization-based scheme. The reward function is formulated with the user sum rate and the SINR against jamming. A transfer learning technique called hotbooting exploits similar communication experiences to initialize the Q-values for faster initial learning. A Dyna architecture is designed to generate hypothetical experiences to update the Q-values for higher learning efficiency. In the

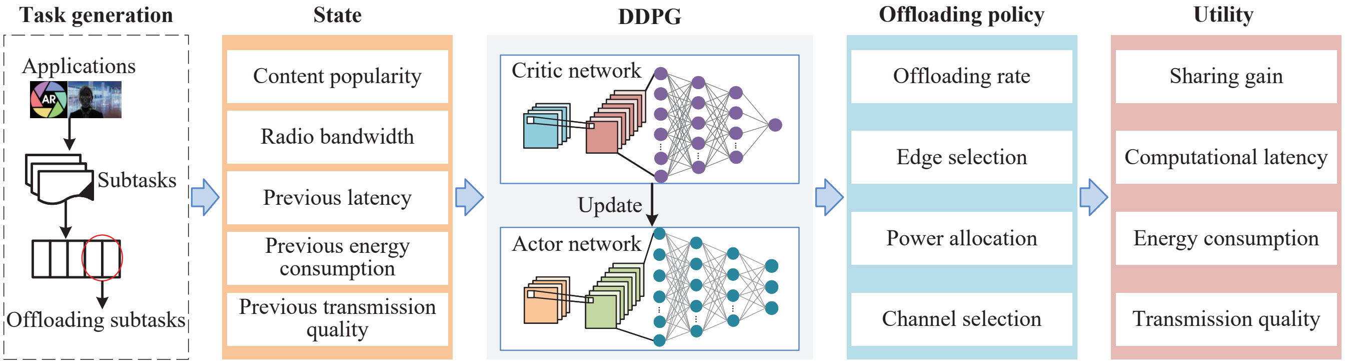


Fig. 4. Deep RL based mobile offloading against jamming and interference.

simulations with 10-antenna BS with transmit power up to 20 W against a jammer that applies Q-learning to choose the jamming power up to 20 W, the RL based schemes exceed the benchmark Q-learning based OMA scheme. For example, the scheme has 10% higher SINR, 12% higher sum rate and 15% higher reward compared with the benchmark for the user with 8 antennas against the smart jammer with 2 antennas. Nevertheless, the anti-jamming performance should be further improved for large-scale NOMA systems with high-dimensional state space.

In another RL based anti-jamming NOMA communication system, a three-level hierarchical DQN helps optimize the BS subband selection, antenna selection and power allocation of the NOMA system Dual-SIC-JAT, as presented in [108], to further improve the user sum rate and learning efficiency and reduce the BS energy consumption in a more complicated system than [42]. More specifically, in the first level DQN, the transmit subband is chosen from the available subbands based on the state consisting of the jamming power and the SINR of the BS signals received by the users. In the second level, the transmit antenna is selected based on the chosen subband and the state. The third level chooses the transmit power coefficient for each user based on the chosen transmit subband and antenna and the state in this time slot. The users with limited computing capacity cannot support the complex DQN, thus having anti-jamming performance degradation in the uplink communication.

B. MEC

The optimal anti-jamming MEC communication policy relies on the jammer location, the jamming policy, the task generation model at the mobile devices, and the jammer channel states, which are rarely obtained by the mobile devices. Therefore, the anti-jamming MEC system in [98] combines safe exploration with Q-learning to choose the edge device, the transmit power and the offloading rate. The state includes the popularity of each subtask, the bandwidth between the mobile device and the edge device, and the previous transmission quality and computational overhead. The reward is formulated with the BER of the messages from the mobile device, the total computational latency and the energy consumption of the mobile device.

The computational latency as an RL security criterion is used to evaluate the risk values in the policy selection to avoid choosing risky policies related to task failure. The anti-jamming offloading game is formulated between the jammer and the mobile device and the Nash equilibrium is derived for the case that the mobile device generates a large number of tasks and has to offload all the tasks to the edge device with a good channel condition. In the simulations, a mobile device generates 300 Kb tasks in every time slot and is connected to three edge devices with 2.5 to 10 MHz radio bandwidth against a Q-learning based jammer with jamming power ranging from zero to 100 mW. The results show that this scheme reduces 23.4% of the computational latency, saves 25.9% energy consumption, improves the reward by 46.5% compared with SEGUE scheme in [114] after 1500 time slots, and converges to the performance bound.

As illustrated in Fig. 4, DDPG is applied for the mobile device with sufficient computational resources to further reduce the task latency and the energy consumption in the continuous action set. Four CNNs (i.e., the actor network, the critic network, and the two target networks) are designed to choose the offloading policy, each having two convolutional (Conv.) layers and four fully connected (FC) layers. More specifically, the state sequence including the current state and the three previous state-action pairs is reshaped into a 93×1 matrix and then input to the actor network that outputs a 3-dimensional vector corresponding to the power allocation, the edge selection and the offloading rate. The state sequence, chosen offloading policy, and reward are used to formulate the offloading experience, which is saved into the experience pool in every time slot. By randomly sampling several offloading experiences, a minibatch is formulated as input to the critic network, which outputs the Q-value of the chosen offloading policy under current state sequence. This scheme further reduces the computational latency to 92.6 ms to satisfy the requirement of the MMORPG game PlaneShift. However, this scheme does not consider the impact of the other mobile devices on the anti-jamming communication performance, which has performance degradation in a cooperative MEC system.

C. IRS

The IRS is equipped with a large number of low-cost passive reflecting elements with flexible reflection amplitude and shift phase. It enhances the desired signal and weakens the undesirable signal to resist jamming [109]. The selection of the optimal number of reflecting elements and the IRS phase shift of each element that determines the transmission quality, such as the BER, depends on the accurate jamming channel states and jamming policies.

The IRS-aided anti-jamming communication system in [109] applies WoLF-PHC and fuzzy state aggregation to choose the BS transmit power and the IRS phase shifts against smart jamming. The system formulates the reward based on the achievable data rate, the BS energy consumption and the outage probability. The transmission policy depends on the state that consists of the received jamming power, the SINR of the signals received by the users and the BS-user channel gain. In the simulations, an eight-antenna BS sends messages with up to 40 dBm transmit power to the four users aided by an IRS with 20 ~ 100 elements against an eight-antenna smart jammer, which moves randomly near the users and applies Q-learning to choose its jamming power ranging from 15 to 40 dBm. The results verify the performance gain of this scheme in terms of the system rate and SINR protection level over the fast Q-learning based anti-jamming scheme. This scheme can have performance degradation and even transmission failure in a more complicated IRS-aided 6G system with a large number of available states and policies.

Intra-agent transfer RL can further improve the learning speed of [109]. For example, the IRS system with reflection beamforming [110] can apply transfer learning based DDPG to jointly optimize the UAV location and the IRS beamforming vector for each element with the goal of maximizing the long-term expected reward including the transmission rate and computational overhead. The state consists of the signal SINR, the IRS channel states, and the jamming power and is input to the DDPG including both an online actor network and an online critic network. Each network has two Conv. layers and two FC layers, whose weights are initialized based on the similar anti-jamming experiences from the IRS controller exploited by transfer learning. This scheme relies on an IRS controller with sufficient energy and computational resources. It can suffer from high computational complexity and latency in the resource-constrained IRS system.

D. MmWave

MmWave systems with a spectrum ranging between 30 GHz and 300 GHz such as the mmWave MIMO system as proposed in [111] can apply RL to optimize the beamforming vector and power allocation without relying on the knowledge of the jamming policies such as the interval and channel states, and the number of jammers. Deep reinforcement learning such as A2C can further improve the anti-jamming performance for the mmWave multiuser system with a high-dimensional state, e.g., the transmitter with 15 mmWave propagation channels against a jammer with two omni-directional antennas. More specifically, the state that consists of the BER of the BS messages and the channel gain vector among the BS and the users is input to both the actor network and the critic network each of which has three FC layers. The actor network outputs the advantage function of the state-action pairs and the critic network outputs the value function of the state. The policy distribution based on both the advantage function and the value function is used to choose the BS transmit power and the beamforming vector from the multi-antenna BS to the multi-antenna users to increase the reward including the average BER of the BS messages, the outage probability and the sum rate. As for implementation, the multi-antenna BS in the mmWave system must have sufficient energy resources to support deep learning and to optimize the anti-jamming communication policy faster than the changing rate of network topology.

E. Massive MIMO

Massive MIMO systems have a large number of antennas and support multiple users in different locations, which result in large variations of received signal strength among different users, and make the users and BSs vulnerable to jamming attacks. Power allocation has been widely used in both the downlink communication [115] and uplink communication [17], [113], [116] for massive MIMO systems to resist the jammers that have the accurate BS and user location information. However, these schemes rely on the perfect knowledge of the channel states of the user-BS and jammer-user/BS links, and thus have performance degradation under the dynamic network topology.

As a novel RL based anti-jamming communication scheme without knowing the jamming model and channel model, the massive MIMO BS [112] applies PHC to optimize the transmit power based on the jamming power and the SINR vector of the signal received by the users to improve the reward that relies on the SINR, the user sum rate and the BS energy consumption. By using transfer learning and data mining technique, this scheme further improves the learning efficiency with the exploited anti-jamming experiences from several simulated scenarios. In the simulations, the BS with 16 radio-frequency chains and 48 ~ 256 transmit antennas sends the messages to 16 single-antenna users with power ranging from 1 ~ 10 W against a smart jammer with a single antenna. Simulation results show that this scheme increases the SINR by 14.0%, the sum rate by 18.0%, and the reward by 40.0% compared with the benchmark power control and rate adaption scheme. Nevertheless, this scheme suffers from a high curse of dimensionality in the massive MIMO system with a large number of users and multi-antenna BSs, which causes a high outage probability.

RL helps improve the achievable rate and save the user energy efficiency of the massive MIMO uplink transmission system in [17] without depending on the location and frequency of the single-antenna jammer. In particular, DQN involving two CNNs each of which has two Conv. layers and two FC layers can accelerate the learning speed for the high-dimensional state space in a system with multi-antenna BSs and a large number of users, e.g., a massive MIMO system with a BS that has more than 400 antennas. Specifically, the state contains the message type (such as the control message transmitted on the control channel), the SINR received at each antenna of the BS, the jamming power and the channel gain vector of the user to all the antennas of the BS. The previous state-action pairs and the current state are used to formulate the state sequence, which is reshaped into the online CNN that generates the Q-values estimated via the neural network function approximators to choose the user transmit power ranging between zero to 5 dB. The target network is updated with given time duration to avoid the unstable power allocation policy explorations. This scheme can improve the immediate reward that is the sum of the achievable rate and energy efficiency, but may suffer from a transmission failure with the delayed feedback after performing the chosen policy.

Multi-agent RL can further improve the anti-jamming performance of the massive MIMO with a large number of users. For example, the MADDPG based massive spatially correlated MIMO system can apply the minimum mean-squared error based jamming suppression scheme as presented in [113], which jointly optimizes the transmit power of each user in both the training phase and the data transmission phase to the BS without the prior knowledge of the jammer-BS channel states. More specifically, each user incorporates its own observation that can be the user-BS channel gain and the user signal SINR at the BS, and the anti-jamming experiences shared by the other users in the system as the state in this time slot, which are input to the actor network that outputs the corresponding transmit power level. The states and power levels of all the users are used to update the global critic network to increase the system reward including the energy consumption and the sum rate. However, this scheme highly relies on the shared experiences from other users and has severe learning efficiency degradation with selfish and malicious users.

In summary, the RL based anti-jamming solutions can improve the communication reliability for NOMA, MEC, IRS and massive MIMO systems. The jamming resistance of a mmWave system can be enhanced by deep RL based on accurate channel estimation, but the performance degrades with partial observations of the network and jamming states, and the delayed feedback of the communication performance from the environment.

V. RL BASED SECURE COMMUNICATIONS

RL enables users and BSs with NOMA, MEC, VLC, THz, IRS, mmWave and massive MIMO to optimize the policy such as the power allocation, beamforming, artificial noise (AN) strategy, relay selection, precoding and the IRS phase shifts. The goal is to improve the secure communication performance such as the secrecy rate and the BER, as summarized in Table IV.

A. NOMA

RL based secure NOMA system can optimize the communication policy (such as the subchannel selection of each user and the BS transmit power) from the N feasible actions without depending on the wiretap channel states and thus improve the communication performance. The reward function can be the sum of the secrecy rate minus the intercept probability that represents the fraction of the successfully intercepted data by an attacker among all the data received at the legitimate receiver, the outage probability and the energy consumption. A typical NOMA system such as [123] has a large number of secondary users with multiple antennas, yielding a high-dimensional state space in the learning process, which significantly degrades the learning efficiency of RL algorithms such as Q learning. Therefore, a four-layer CNN connecting to N DNDs further improves the learning speed similar to NEC. More specifically, the state consisting of the transmit power of the primary BSs and users and the received signal strength indicator (RSSI) at all the secondary users is compressed with the pooling layers and input to the CNN. The CNN has network weights initialized with transfer learning techniques such as hotbooting in [42] and outputs the DND keys, which are input to each DND. Each DND outputs a Q-value corresponding to a transmission policy under the state in this time slot to increase the NOMA reward, but cannot avoid the vulnerable communication policies that degrade the message reception at the BS and even cause user-BS transmission failures.

Another RL based secure NOMA system can apply a deep RL algorithm such as safe DQN to choose the BS transmit power and the AN beamforming vector to further increase the reward based on the secrecy rate, the eavesdropping rate and the energy consumption of [123]. To satisfy the user QoS requirement (e.g., the transmission rate for all the users in [18]), a security criterion can be designed based on the QoS requirement to avoid exploring the dangerous communication policies that result in severe data leakage. For example, the risk value is set as the indicator function that represents whether the sum secrecy rate is less than the QoS requirement. The state contains the estimated wiretap channel state, the channel state information among users, the system energy efficiency and the inter-user interference signal strength. The state is input to both the Q-network and the E-network each consisting of three FC layers, i.e., an input layer with a size equal to the state dimensions, a hidden layer having a size based on the learning samples and state dimensions, and an output layer with a size equal to the number of available transmission policies. The Q-network outputs the Q-values of the transmission policy of the secure users, while the E-network outputs the long-term risk values under the current state. Both the Q-values and the long-term risk values formulate the Boltzmann policy distribution to improve the secure performance of the NOMA system.

B. MEC

MEC can apply convex optimization and unsupervised learning to optimize the secure communication policies such as the transmit power and the mobility policy [124]–[127], the offloading policy [78], [128], and the edge computational resource allocation [129]. The optimal MEC secure communication policies rely on the eavesdropper location and policy and the wiretap channel states, which are rarely obtained by the mobile devices and the edge devices. Therefore, the RL

TABLE IV
RL BASED SECURE COMMUNICATIONS

Learning agent	**Policy**	**Reward**	**RL algorithms**	**Applications**
MEC server [117]	Offloading policy Blockchain strategy Computing resource Radio bandwidth allocation	Smart contract fee Computation latency Energy consumption	DDPG	MEC
Transmitter [32]	Beamforming vector	BER Secrecy rate	Q-learning DDPG	VLC
BS [118]	Power allocation IRS phase shifts Beamforming vector	Achievable rate Secrecy rate Energy consumption Outage probability QoS requirements	PDS DQN	IRS
Relay node [119]	IRS phase shifts	Secrecy rate Throughput	A3C	
BS [120]	Beam width Power allocation User association	Achievable rate Energy consumption	Risk-sensitive RL	MmWave
Vehicle [121]	Power allocation Communication mode	Transmission latency Network throughput	DDQN	
BS [122]	Precoding matrix AN shaping matrix Power allocation	Packet delay Secrecy rate	A3C	Massive MIMO

based MEC secure communication schemes improve the performance such as the secrecy rate and the intercept probability against eavesdropping [117].

The blockchain-aided secure MEC system in [117] applies DDPG involving an actor network and a critic network to determine the offloading policy of the L mobile devices, the resource allocation of the edge device and the smart contract (i.e., a self-operating computer program running on the blockchain platform) without relying on the eavesdropping pattern. The state including the transmission data rate, the task data size, the total number of CPU cycles and the rest bandwidth resources of the edge device is input to the actor network with three hidden layers, which directly outputs the offloading policy, the computing resource and radio bandwidth allocation, and the blockchain strategy. The goal of the edge device with sufficient computational resources is to maximize the long-term expected reward based on the smart contract fee, computation latency, and energy consumption against an eavesdropper at a fixed location. In a MEC network that contains an edge server with 15 MB bandwidth resource and 2 ~ 30 mobile devices each of which generates 2 MB tasks in each time slot, this scheme reduces the offloading cost by 21.9% compared with the DQN-based offloading scheme.

The deep RL based UAV-enabled secure MEC systems can improve the secrecy capacity without relying on the perfect knowledge of the eavesdropper locations, as compared with [117]. For example, the MEC system can apply DQN to choose both the transmission policy (e.g., the transmit power and the AN strategy) and the offloading policy such as the edge selection and the offloading data size to maximize the long-term expected reward that consists of the secrecy capacity, the energy efficiency, the computational latency and the transmission latency. To address the performance degradation due to the policy quantization under a large-scale network (e.g., 100 mobile devices in a 400 × 400 m^2 square area in [124]), the edge device uses DDPG involving an actor network and a critic network, each with four FC layers to reduce the computational complexity, to improve the MEC secure communication performance. Intra-agent transfer learning can be used to exploit secure experiences to initialize the weights of the actor network, and a critic network is used to update the weights to accelerate the initial policy learning process. The state contains the battery level, the estimated eavesdropping rate, the radio bandwidth to the mobile devices, the energy consumption and the computational latency. With the state as the input, the first layer of the actor network chooses the secure MEC policy that consists of the edge device location, the friendly jamming power, the transmit power and the offloading rate. However, this potential scheme is based on how to quantize the edge device location, friendly jamming power, and transmit power for smaller quantization errors as well as shorter exploration time in the learning process.

To further improve the secure communication performance of the large-scale MEC as in [78], deep RL such as PPO can be applied to optimize the friendly jamming power and offloading policy. The reward function depends on the secrecy capacity, the energy consumption and the computational latency. Both the actor and critic networks are also used in this scheme, which have the same architecture including an input layer, a hidden layer and an output layer to replace the Conv. layers for less computational complexity. The state includes the eavesdropping channel state, the computational latency,

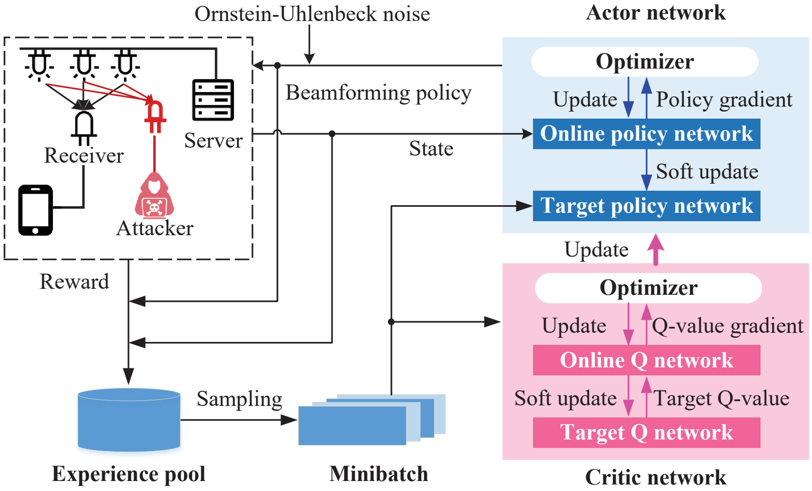


Fig. 5. Deep RL based VLC secure communications.

and the energy consumption. With the state as the input, the actor network outputs the policy distribution based on the multivariate Gaussian distribution, while the critic network estimates the state values of the policy to update the actor network weights.

C. VLC

The friendly jamming based VLC secure communication scheme in [130] determines the optical jamming policy based on the channel gain of the source-eavesdropper link that is rarely known in advance. Therefore, the RL based VLC secure communication system in [32] chooses the beamforming policy for the legitimate user without relying on the known wiretap channel states against a passive eavesdropper at a fixed location. The scheme applies both Q-learning and DDPG to optimize the secure communication policy based on the state that consists of the channel state information of the transmitter, the BER of the messages and the secrecy rate to increase the reward, including the secrecy rate and the BER.

The DDPG architecture with two networks is shown in Fig. 5. The actor network that consists of two Conv. layers and two FC layers with the state as the input outputs the secure communication policy (i.e., the beamforming vector) with the Ornstein-Uhlenbeck noise. The transmitter chooses a learning rate that is much smaller than one to soft update the weights of the target actor network by copying the weights of the online actor network. It ensures that the output of the target actor network changes slowly, thus improving the learning stability and robustness. The critic network with the sampled experiences from the minibatch as the input estimates the chosen policy to update the weights of the actor network. Experiments are performed based on a $5 \times 5 \times 3$ m^3 room equipped with 36 light-emitting diodes, in which a transmitter located at 85 cm height sends the streaming content to a receiver located at 3 m height with 60° receiver field-of-view against a passive eavesdropper located at 85 cm hight verify the performance gain over the benchmark scheme. For example, the scheme improves the secrecy rate by 116.3% to 2.033 bps/Hz, decreases the BER by 85.5% to 4.5‰ and increases the reward by 2.39 times compared with the fixed friendly jamming scheme after 5000 time slots. The deep RL version further improves the secrecy rate by 29.7%, decreases the BER by 79.3% and increases the reward by 29.8%.

In the deep RL based secure multiple input single output VLC system similar to [131], a two-level hierarchical DQN is designed for the transmitter with sufficient energy and computational resources to optimize the beamforming vector and jamming power of the user with J LED transmitters to increase the reward, including the peak SINR of the legitimate signals and the secrecy rate. More specifically, the state containing the channel gain vector from the transmitters to the receiver, the peak SINR at both the legitimate user and the passive eavesdropper and the secrecy capacity is input to the first level. Consisting of three FC layers, the first level of DQN outputs J Q-values to choose the beamforming vector. The second level selects the friendly jamming power based on the state and the chosen beamforming vector.

D. THz

The secure THz communication systems [132]–[135] require the full knowledge of the eavesdropping probability, the user distribution and density, the labelled path data and the eavesdropper location. Therefore, the THz communication system can apply RL such as Dyna-Q, PDS and PPO to optimize the transmission parameters such as the transmit power without knowing the wiretap channel states against both passive and active eavesdropping. For example, the state is formulated with the evaluated secrecy rate, the measured BER and the estimated channel states of the receivers. The state is used in the optimization of the propagation paths and the transmit power in terms of the eavesdropping rate and the energy consumption.

The RL based secure THz communication helps improve the secrecy capacity of [133] that exploits the unique spectrum features of frequency-dependent molecular absorption against passive eavesdropping. In this scheme, the state that consists

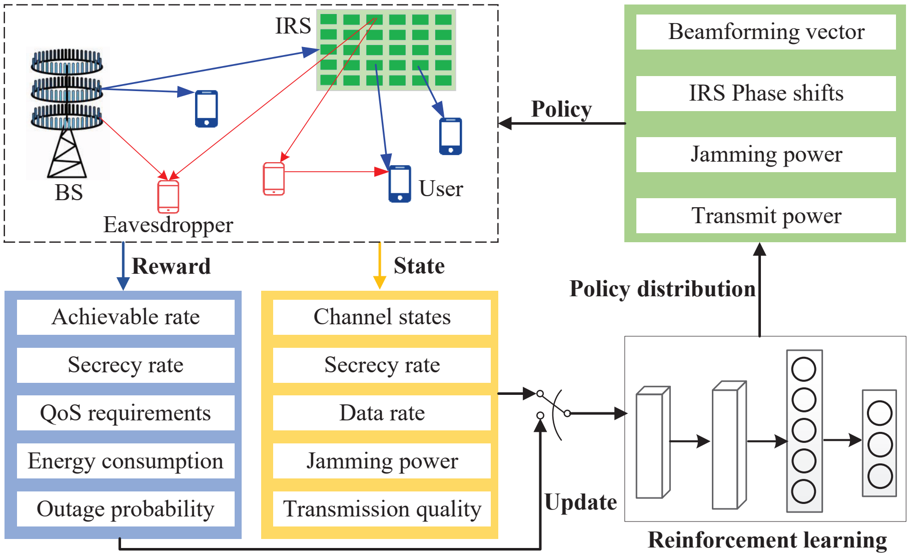


Fig. 6. RL based IRS-aided communications against active eavesdroppers.

of the SINR at the receiver, the maximal covert data rate and the estimated intercept probability is used as the input of A2C with an actor network and a critic network each of which has three FC layers. More specifically, the actor network estimates the advantage function of the baseband waveform, the transmit power and the carrier frequency, and the critic network evaluates the current value function that relies on the eavesdropping rate and the secrecy outage probability. The policy distribution is formulated based on both the advantage function and the value function to avoid local optimum policies in the learning process.

E. IRS

The reflecting beamforming coefficient that determines the secrecy rate relies on the channel state information of the attackers and the attack frequency or pattern [136]–[138]. For example, the secure communication scheme in [137] applies alternating optimization and semidefinite relaxation to jointly optimize the AP transmit beamforming and the IRS reflect beamforming against a single-antenna eavesdropper. Nevertheless, this scheme assumes the quasi-static flat-fading channel model and thus suffers from performance degradation in dynamic and complicated networks. Therefore, an RL based IRS-assisted MIMO secure communication scheme is presented in [118] to further improve the secrecy rate in dynamic systems.

This scheme combines PDS with DQN involving a multi-layer perceptron network to optimize the IRS phase shifts, the BS transmit power and the beamforming vector based on the state that consists of the channel states of all users, the secrecy rate, the data rate, the jamming power and the transmission quality against passive eavesdropping. The achievable rate, the secrecy rate, the energy consumption, the outage probability and the QoS requirements are used to form the reward function, as illustrated in Fig. 6. Simulations are performed in the system with a 4-antenna BS with power ranging from 15 to 40 dBm aided by an IRS with 10 ∼ 60 elements and two single-antenna mobile users based on the log-distance path loss model against two single-antenna passive eavesdroppers. The results show that this scheme improves the secrecy rate by 17.2% and the QoS satisfaction level by 8.7% with 0.7 channel state information coefficient of the two mobile users compared with the DQN-based beamforming scheme. However, this scheme may suffer from severe communication performance degradation due to the high-dimensional state space and the significant increase in the available policies for the IRS-aided system with a large number of users.

The multi-agent RL based IRS-aided secure communication in [119] applies distributed A3C to choose the IRS phase shifts for the source and the transmission link for each relay and improves the learning performance of the single-agent RL based scheme in [118]. The state that consists of the buffer states, the channel states of the relay nodes and the channel state information among the relay nodes, source and IRS is used in the optimization of the secrecy rate and throughput following a delay constraint. In a secure cooperative network aided by an IRS with 32 elements that consists of five relays against a passive eavesdropper, this scheme improves the average secrecy rate by 60.0% and the throughput by 65.0% compared with the max-ratio scheme. This scheme highly relies on the shared experiences from the other relay nodes and thus may waste a large number of random exploration time slots, if the system has malicious or selfish nodes.

The intra-agent transfer RL based secure IRS-aided MIMO system applies transfer learning to accelerate the learning speed of [119], improves the secrecy rate, and reduces the BS energy consumption of the alternating optimization based scheme in [139]. At the beginning of the learning process, the BS exploits the secure communication experiences in the previous tasks to initialize the learning parameters of deep RL, such as the network weights, the learning rate and the discount factor. The BS transmit power, the AN strategy and the IRS phase shifts are chosen with a four-layer PPO based on the state that contains the outage probability and the intercept probability rather than the wiretap channel states.

A three-level hierarchical DQN can further improve the secure communication performance of the IRS-aided MIMO system in [140] with a large number of users. More specifically, the optimal transmit beamforming vector, AN covariance matrix at the AP and IRS phase shifts maximizes the long-term expected reward that depends on the sum rate, the eavesdropping rate, the intercept probability and the AP energy consumption. The state that includes the received signal strength at the legitimate users, the outage probability and the eavesdropping data size is input to the first level DQN that outputs the Q-values of the feasible power levels to choose the transmit power. The second level chooses the jamming power to send AN signals to the passive eavesdropper based on the state and the chosen transmit power. In the third level, the policy distribution of the IRS phase shifts is formulated based on the state and the chosen transmit power and jamming power from the two higher levels.

A DDPG based IRS-aided system is designed to jointly optimize the BS transmit power, AN covariance matrix, and IRS phase shifts of the MIMO system in [141] to reduce the quantization errors of the action set without relying on the wiretap channel states and the eavesdropper location. The secure communication policy is chosen by the online actor network of DDPG based on the state that consists of the channel gain vector among the BS, IRS and users, the data priority and the eavesdropping rate to improve the reward, including the secrecy rate, the BS energy consumption and the outage probability. In the network update process, the online critic network updates the weights of the online actor network with the evaluated Q-values, and two target DNNs are soft updated for more stable exploration under the correlated states.

F. MmWave

The mmWave secure communication policy includes the precoding [142], [143], beamforming [144], [145], channel frequency selection [146], relay selection [147], power allocation [148] and friendly jamming [147]. It relies on the perfect knowledge of the attacker mode or pattern that is rarely obtained by practical BSs, and thus the secrecy rate sharply decreases under active eavesdropping that combines sniffing with jamming. Therefore, an RL based downlink mmWave system in [120] applies a distributed risk-sensitive reinforcement learning to choose the beam width, transmit power, and user association under the discrete action set and state space. More specifically, the state consists of the network queuing status and the downlink channel states. The reward depends on the BS energy consumption and the achievable rate of the associated users. In the simulations based on the 64×4 downlink transmission with 24 small cells, this scheme improves the network reliability by 11.1%, increases the availability by 20.0% and accelerates the policy optimization speed by 5 times compared with the Q-learning based scheme in a range of 0.5×0.5 km^2.

The mmWave vehicular communication in [121] applies DDQN that consists of an online network and a target network with the same architecture to choose the transmit power, and the communication mode that includes the cellular mode, the dedicated mode and the reuse mode from all the available policies under each potential state without dangerous exploration avoidance. The state consists of the vehicle-to-vehicle channel states, the vehicle-to-infrastructure interference, the neighboring channel selection, the transmission load and the QoS requirement. The reward relies on the transmission latency and the network throughput. The online network has an input layer with 82 neurons, three hidden layers and an output layer with 9 neurons. In a vehicular network with 30 resource blocks, this scheme decreases the sum throughput from 165 to 148 bps as the vehicle speed changes from 10 to 60 km/h, which is 55.8% higher than the benchmark.

In the secure mmWave beamforming system with a frequency diverse array as presented in [144], CNNs combined with DND can be used to optimize the frequency offset increment vectors of the transmitter. The state can be the channel gain vector of the transmitter-receiver link, the secrecy rate requirement and the SINR against a sensitive eavesdropper that can intercept user data via a sidelobe. The reward includes the secrecy outage probability, the intercept probability and the energy consumption. The state sequence rather than the state is input to the CNN with two Conv. layers and two FC layers, and the DNDs output the Q-values of the available policies in the action set.

The RL based secure mmWave communication can apply A2C to choose the codeword rate and the transmit power of both the source and the relay to improve the secrecy throughput and save the system energy consumption of the convex optimization based scheme presented in [147]. The state that consists of the data priority, the source-relay distance, the signal SINR at both the relay and the destination and the intercept probability is input to the actor network of A2C involving three FC layers. With the same network architecture as the actor network, the critic network evaluates the long-term expected reward of the chosen transmission policy under the state in this time slot.

G. Massive MIMO

In massive MIMO systems, secure communication based on AN or friendly jamming applies convex optimization to choose both the source and relay transmit power levels [149]–[151], the transmission duration [150], and the downlink precoding [152]. For example, the secure massive MIMO communication in [152] assumes the accurate wiretap channel states against active eavesdropping that combines pilot spoofing attack and uplink jamming. However, this system has severe performance degradation under the unknown eavesdropping probability, location and channel states. Therefore, a RL based secure massive MIMO system is proposed in [122] without depending on the known attacker location.

In this system, A3C with a policy network and a value network each with two FC layers is applied to choose the precoding matrix, the AN shaping matrix and the power allocation to users. The state consists of the estimated eavesdropping channel states and the uplink channel matrix. The reward

TABLE V
RL BASED PHY-LAYER AUTHENTICATION

Learning agent	Policy	Reward	RL algorithms	Attacks
Estimated attack rate Precision rate Recall rate	Test threshold	Detection accuracy Communication overhead	DDPG [153]	Spoofing attacks
Number of users Detection latency Miss detection rate False alarm rate	Spreading codes	Detection accuracy	DDQN [77]	Spoofing attacks
Authentication accuracy Attack rate Message priority	Authentication mode Test threshold	Authentication accuracy	Dyna-Q NEC [44]	Spoofing attacks
Attack features Types of the attackers Attack strength	Edge detection mode	Detection accuracy Communication delay	DQN [154]	DDoS Sybil attacks
Signal coverage probability Message type Number of messages	AP deployment	False alarm rate Miss detection rate Authentication latency	DQN [155]	DoS Injection attacks Spoofing attacks Man-in-the-middle attacks
Received signal strength User location Estimated attack rate Authentication accuracy	Test threshold	Detection accuracy	PPO [156]	Spoofing attacks Injection attacks

function is the sum secrecy rate and cumulated packet delay, while neglecting the user energy consumption. Simulation results in a 256×1 MIMO system with four users each of which has 40 dBm transmit power and 10 MHz bandwidth against a single-antenna eavesdropper show that the scheme reduces the cumulated packet delay by about 25.0% compared with the benchmark randomized policy scheme.

The RL based massive MIMO secure communication with decode-and-forward applies safe Q-learning to optimize the relay power selection to increase the secrecy capacity of [150]. More specifically, the state that contains the BER of messages, the channel state vector between the relay and the destination, and the eavesdropping rate is used as the basis to update the Q-values. The secrecy outage capacity is used as a security criterion to evaluate the risk level of the chosen relay power, which is used to update the E-values. Instead of using an ϵ-greedy based policy distribution, both the Q-values and the E-values are used to formulate a Bolztman policy distribution to avoid dangerous relay power levels that result in severe data leakage.

Multi-agent RL such as MADDPG improves the secure communication policy optimization efficiency of the AN-assisted massive MIMO systems with a large number of multi-antenna APs and users (e.g., 150 APs in [149]). The transmit power of each AP is chosen to improve the ergodic secrecy rate, reduce the BER of messages and save the AP energy against the multi-antenna active eavesdroppers. The observation of each AP consists of the uplink channel gain vector among users, the number of neighboring APs, the data priority and the jamming signal strength, which is input to the actor network. More specifically, each AP has an actor network that relies on its own observation and directly outputs the transmit power. On the other hand, a global critic network based on the observations and chosen transmit power levels of all the APs evaluates the corresponding Q-values and thus updates the weights of each actor network.

In summary, deep RL algorithms such as DQN, A3C and DDPG have been used to improve the secure communication performance of MEC, VLC, IRS, mmWave and massive MIMO systems. Besides, the anti-eavesdropping performance such as the secrecy rate of NOMA and THz communication can be improved by the tabular RL algorithms such as Dyna-Q and PDS, the policy gradient RL including PPO and A2C and the multi-agent RL such as MADDPG without the full knowledge of the attack patterns.

VI. RL BASED PHY-LAYER AUTHENTICATION

Physical features of the 6G signals such as the channel states and the signal phase offsets can be exploited as the fingerprints of the radio transmitter and reduce the communication and computational overhead of the 6G authentication compared with the authentication solely relying on cryptography, trust and certificate. As summarized in Table V, RL enables the wireless devices in the communication systems with NOMA, MEC, mmWave and massive MIMO to optimize the authentication/detection mode and parameters, the offloading policy, the resource allocation, and the block size and interval to resist spoofing attacks without knowing the accurate attack frequency or location [44].

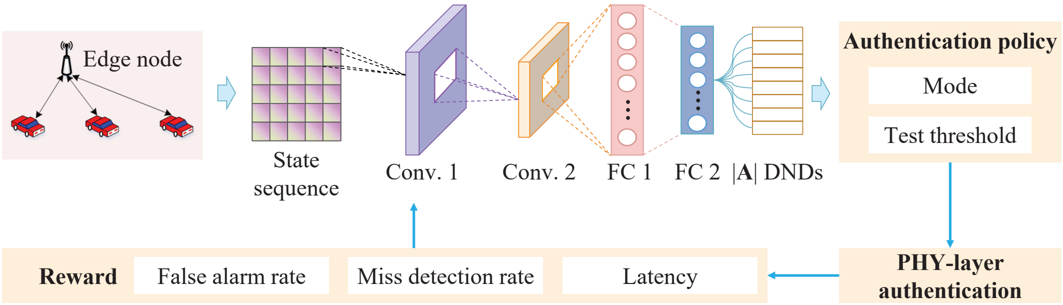


Fig. 7. Deep RL based PHY-layer authentication in MEC systems.

A. NOMA

Deep RL such as DQN can help choose the test threshold for the NOMA systems with high-dimensional state space (e.g., the 256×256 system with 100 channel wavelengths as presented in [153]) without knowing the time interval of spoofing messages. Specifically, the NOMA system uses the current state including the estimated attack rate, the precision rate and the recall rate as the input of DQN having an input layer with three neurons, a hidden layer with 64 neurons and an output layer with size depending on the quantized test threshold levels. The DQN outputs the Q-values for each state-action pair in the current time slot, which are used to choose the test threshold according to the ϵ-greedy algorithm. Furthermore, DDPG can reduce the quantization error of the test threshold in a continuous action set, thus improving the detection performance of DQN. More specifically, the BS inputs the state that includes the estimated attack rate, the precision rate and the recall rate to the online actor network of DDPG that outputs the test threshold ranging from zero to 1. The authentication experience including the current state, the next state, the chosen test threshold and the reward that involves the detection accuracy and the communication overhead is used as the input of the critic network that updates the weights of the actor network. Nevertheless, this potential scheme uses the characteristic of the massive MIMO mmWave channel to achieve principal component analysis detection, thus suffering from low accuracy under lightweight IoT systems that cannot simultaneously support mmWave and massive MIMO techniques.

The DDQN based spoofing detection enables the BS in the NOMA system as presented in [77] to choose the number of spreading codes for each user against the spoofing attackers that can send the same pilot sequence as the legitimate user. More specifically, the previous B states each of which consists of the number of users, the detection latency, the miss detection rate and the false alarm rate are used to formulate the state sequence, which is reshaped into the online CNN with two Conv. layers and four FC layers. The target network evaluates the estimated Q-values that rely on the reconstruction and detection accuracy to avoid the over-estimation error of the online network. This scheme aims to improve the reward including the detection accuracy and latency for the single-antenna users.

B. MEC

The RL based authentication system in [44] applies Dyna-Q to optimize both the authentication mode or feature and the authentication threshold in the PHY-layer based hypothesis test, in which the mobile device has $|\mathbf{A}|$ feasible authentication policies. The state is formulated based on the quantized authentication accuracy averaged over 200 time slots from the feedback channel, the attack rate estimated by the mobile device in the legitimate vehicle and the message priority. The mobile device with sufficient computational resources to execute deep learning combines NEC with the intra-agent transfer learning techniques such as the hotbooting to further improve the authentication accuracy. As illustrated in Fig. 7, the NEC consisting two Conv. layers, two FC layers and DNDs outputs the policy distribution of the $|\mathbf{A}|$ feasible authentication policies. Experiments based on a vehicle moving at a speed of 30 km/h carrying both the mobile device and the edge device surrounded by five radio devices with power changing from 50 mW to 3 W, and a rogue edge device moving at 36 km/h with transmit power 100 mW verify the efficacy of the scheme. For example, the false alarm rate and the miss detection rate are decreased by 52.3% and 79.4% respectively compared with the Q-learning based PHY-layer authentication. The system has higher accuracy with the mobile device and the edge device in the same vehicle, but has lower accuracy and higher latency if the edge device locates outside the vehicle that carries the mobile device.

The RL based vehicular authentication system in [154] applies DQN that defines the loss function as the temporal difference error of the Q-value between two successive iterations to choose the edge detection mode against DDoS, Sybil attacks and rogue APs in vehicular ad hoc networks. The state consists of the attack features, the types of attackers and the attack strength, and the reward depends on the detection accuracy and the communication delay. The detection based on the service function chain with the channel monitoring, attacking signature abstraction, signature matching and signature normalization verifies the performance gain over the greedy algorithm by reducing about 80.0% computational cost. This scheme can seamlessly integrate the signature-based and feature-based intrusion detection methods and support various communication types, including vehicle-to-infrastructure, vehicle-to-vehicle and vehicle-to-everything detection scenarios. It may have slow learning speed and even overestimation of the Q-values in a more dynamic and heterogeneous vehicular network.

C. Mmwave

In the RL based mmWave authentication that extracts the spatial-temporal information of the beam pattern as the radio fingerprints similar to [155], transfer learning based DQN optimizes the AP deployment, in order to increase the reward, including the false alarm rate, the miss detection rate and the authentication latency. The state includes the signal coverage probability, the message type and the number of messages that are falsely accepted/rejected by the user. More specifically, a four-layer DNN architecture is designed to output the Q-values of the feasible AP deployment policies, which are used to formulate the policy distribution. In the initial learning process, the user uses transfer learning to exploit the authentication experiences from similar scenarios and thus initializes the DNN parameters such as the network weights and the learning rate. However, an accurate quantization method for all the possible AP deployments should be in place before implementing this scheme in practical mmWave systems, as the performance degrades with policy quantization errors.

D. Massive MIMO

The RL based PHY-layer authentication that uses the channel amplitude and the transmitter hardware impairments reduces both the miss detection rate and the false alarm rate of the massive MIMO system in [156] against a spoofer, which attempts to steal user data or to inject some fake aggressive information into the network with the MAC address of the legitimate user. Specifically, a three-layer DNN based PPO optimizes the test threshold based on the state that includes the received signal strength of the BS at each antenna, the user location, the estimated attack rate and the authentication accuracy. This scheme updates the PPO network weights based on the importance sampling technique with the previous experiences rather than the experience replay technique for less sample complexity. Nevertheless, its authentication accuracy degrades under the smart spoofer with multiple antennas that can choose its spoofing strategies by predicting the ongoing defense policy of the massive MIMO system.

In summary, the RL based PHY-layer authentication reduces the false alarm rate, miss detection rate, and communication latency of MEC systems under the discrete action set and state space. As for future direction, deep RL such as DQN and PPO can be used to help improve the authentication accuracy as well as the communication efficiency for NOMA, mmWave and massive MIMO systems against smart spoofing attacks.

VII. RL based Data Privacy-Aware Secure Communications

The privacy-aware communication systems can use RL to choose the device mobility, the offloading policy and the privacy budget without relying on the knowledge of the attacker location and the attack frequency, in order to resist the background knowledge attacks, eavesdropping, inference attacks and differential attacks for 6G systems with NOMA, MEC, mmWave and massive MIMO. Recently, identity-based authentication [168], [169], differential privacy (DP) [157], [170], [171] and encryption [172] depend on trusted third-party auditors to protect privacy, but the communication and computational overheads are too high for the 6G devices with restricted computing and energy resources. This challenge can be addressed by the RL based data privacy-aware secure communications [27], [158]–[160], in which the value-based RL such as PDS and DQN chooses the privacy policy in a discrete action set, while the policy gradient RL (e.g., PPO and DDPG) deals with the continuous action set, as summarized in Table VI. The policy, state and reward for the RL based data privacy protection for 6G systems are illustrated in Fig. 8.

A. NOMA

NOMA communication systems can apply RL algorithms to protect legitimate data from stealing by the adversary that can eavesdrop on the transmission channel as well as intercept the transmitted messages. For instance, the NOMA-enabled industrial IoT system in [162] combines hierarchical federated learning with deep RL to protect user data privacy without the knowledge of communication models among the IoT devices, the edge servers and the cloud server. In this system, the typical DDPG is applied in IoT devices to optimize the transmit power for uploading the learning parameters, allocation of computing resources to the learning tasks, and orchestration policy based on the state instead of the data leakage model. The state including the channel states of the legitimate links, orchestration policies of other IoT devices, and computing capability, is input to the actor network. Both the actor and critic networks consist of five FC layers, in which the actor network directly outputs the privacy policy and the critic network outputs the target Q-values to evaluate the chosen policy. This scheme aims to maximize the long-term expected reward that depends on the latency, energy consumption, transmission power, and computing capability of IoT devices. In the simulations based on three cells each having three IoT devices with up to 10 GHz computing capacity and 1 W transmit power, this scheme reduces about 17.9% energy consumption than the benchmark, but has a low model accuracy with hundreds of IoT devices.

Deep RL such as DDPG is applied in the NOMA system in [163] to optimize the authentication decision at the receiver such as the test threshold from the continuous action set ranging from zero to one. This scheme uses the authentication tag as well as the channel responses from the users to the BSs to improve both the user privacy protection performance and the authentication accuracy. The state includes the channel responses, the number of users, the distance vector from the users to the BS, the arrival interval of the underlying message, the RSSI and the previous authentication performance. It is input to DDPG with an actor network choosing the authentication policy and a critic network updating the network weights, in which each network has the same architecture, i.e., three FC layers. This scheme aims to improve the robustness,

TABLE VI
RL BASED DATA PRIVACY-AWARE SECURE COMMUNICATIONS

Learning agent	**Policy**	**Reward**	**RL algorithms**	**Attacks**
IoT device [27]	Offloading rate Local processing rate	Privacy level Queuing cost Computational overhead	PDS Dyna-Q	Inference attacks
Edge device [55]	Offloading policy Resource allocation	Edge profits Attack rate Latency Energy consumption	Safe DQN Q-learning	Selfish edge attacks Fake service record attacks
Smart edge [157]	Laplace noise	Accuracy Data utility Privacy level	PPO	Inference attacks
Movie recommendation system [158]	Privacy budget	Privacy protection level QoS requirement	DDQN	Inference attacks
Mobile terminal [159]	Moving strategy	Geographical fairness Energy consumption	DDPG	Selfish attacks DoS attacks DDoS attacks
IoT device [160]	Payment strategy	Cost of participant	DQN Q-learning	Selfish attacks
Mobile device [161]	Offloading policy Power allocation Block size and interval	MEC computation rate Transaction throughput	A3C	Eavesdropping
IoT device [162]	Transmit power Resource allocation Orchestration policy	Latency Energy consumption Ttransmission power Computing capability	DDPG	Eavesdropping
BS [163]	Authentication decision	Privacy protection level Authentication accuracy	DDPG	Eavesdropping
User [164]	Authentication threshold	Propagation delay Privacy level	DDQN	Inference attacks
User [165]	Matching parameters	Classification accuracy False positive rate Authentication accuracy Privacy loss level Attack success rate	PPO	Spoofing attacks
BS [166]	Beamforming policy	Privacy protection level QoS requirement	DDPG	Eavesdropping
Server [167]	Transmission strategy	Model classification Accuracy Data privacy level Transmission latency	DQN	Eavesdropping

compatibility, privacy, and security performance. It relies on the accurate observation and immediate feedback from the environment in each time slot.

B. MEC

Value-based RL enables mobile devices or IoT devices to choose the edge device, the payment strategy, and the computational resource from the discrete action set for data protection in MEC systems. For example, the MEC based healthcare data privacy protection scheme as proposed in [27] designs a model based RL algorithm to optimize the offloading rate and the local processing rate from a quantized action set against inference attacks. The state consists of the newly generated healthcare data size, the data priority, the channel state, the harvested energy, the battery level and the amount of data in the buffer. By combining PDS, Dyna architecture with transfer learning, this scheme evaluates the reward based on the privacy level, the queuing cost, and the communication and computational overhead. For the IoT device that generates 30 Kb of healthcare data each second, the scheme improves the privacy level by 36.6%, reduces the latency by 68.8% and the energy consumption by 9.6%, and saves the convergence time by 40.0% compared with the CMDP based offloading scheme after 2200 time slots. The performance effectiveness may degrade under a MEC system with high-dimensional state space, which even results in the curse of dimensionality, thus

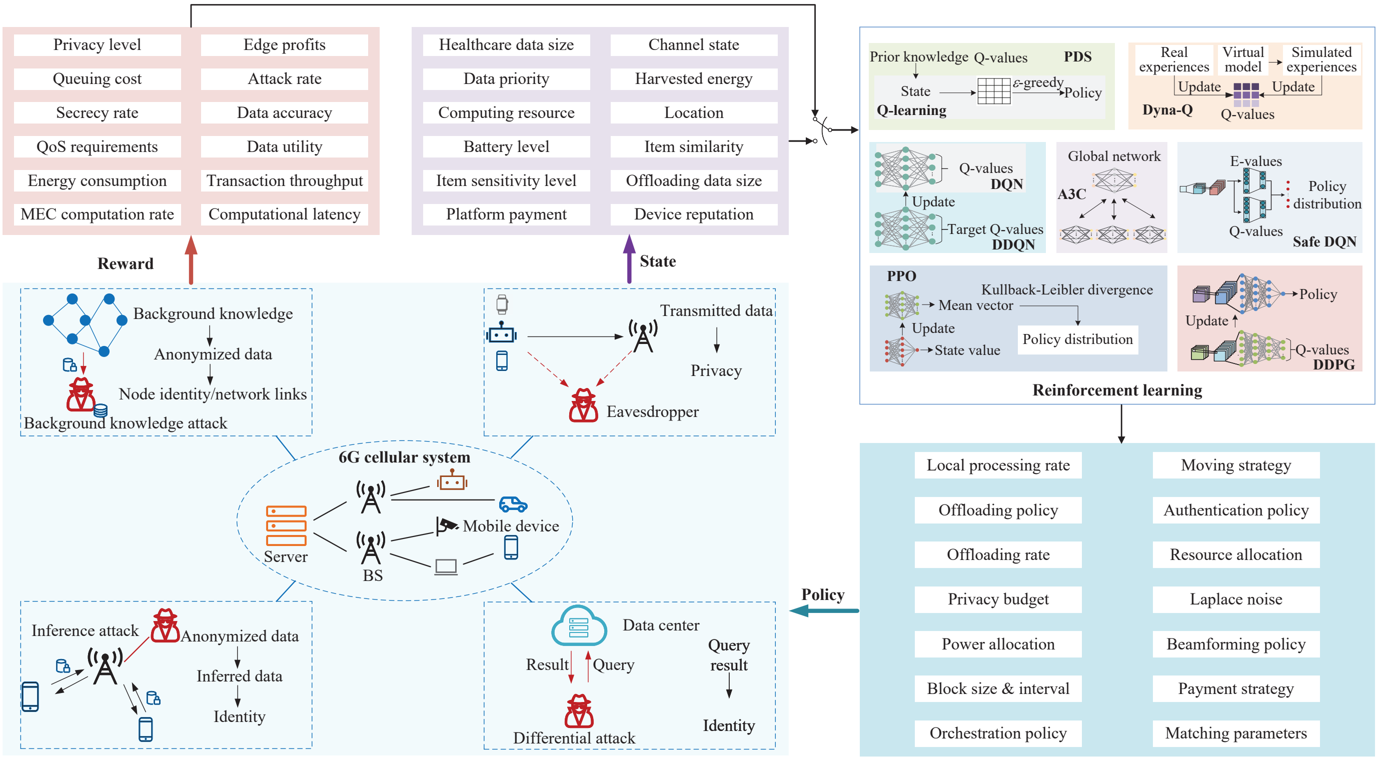


Fig. 8. RL based data privacy protection for 6G systems.

having an extremely slow learning speed.

To further accelerate the policy optimization speed in [27], the IoT privacy-aware communication system in [160] applies DQN with two Conv. layers and two FC layers at the MCS platform to choose the payment strategy based on the first state that consists of the privacy protection levels of the available IoT devices, and applies Q-learning at the IoT devices to select their privacy budgets of the sensing data based on the second state, including the previous privacy level and the platform payment. This scheme improves the privacy protection level and reduces the aggregated error of the platform in a data aggregation crowdsensing system with 60 ∼ 300 IoT devices against selfish attacks. For example, this scheme decreases the average aggregated error of the platform by 12.9% and thus increases the reward of the platform by 15.4% compared with the random strategy scheme at the 5000-th time slot. However, it has to deal with the quantization error issues of the action set and the vulnerable exploration during the whole learning process.

The blockchain-assisted MEC system in [55] applies RL to optimize the offloading policy and computational resource allocation against selfish edge attacks and fake service record attacks. It uses a risk value function to avoid choosing the risky policies related severe privacy leakage compared with [27]. More specifically, the transfer learning based Q-learning is applied to choose the number of edge CPUs to compute the tasks from each mobile device based on the offloading data size and the reputation of each mobile device. This scheme uses the historical reputation vector and the computational latency of each edge device to calculate the current reputation and applies blockchain to record the reputation of the available edge devices. The proof of work (PoW) mechanism is used as the basis to choose the miner in the blockchain. Safe DQN can further increase the reward based on the edge profits, the attack rate, and the task computational performance, as illustrated in Fig. 9. More specifically, the edge device compares the SINR of the signals sent by the mobile device with the QoS requirement and sets the risk value as one if the SINR is smaller than the QoS requirement and zero otherwise. This scheme uses the current and previous risk values to evaluate the E-values similar to [95] and thus enables safe exploration during the offloading process. In the simulations, three mobile devices generate the computational tasks at 10 Mbps, process the local tasks at 200 Kbps, and are connected to three edge devices each with 10 CPUs over the channel with up to 10 Mbps bandwidth. This scheme reduces the response latency by 9.5%, saves the energy consumption of the edge devices by 9.1% and increases the reward by 35.1% compared with the benchmark scheme after 500-th time slot. However, this scheme does not account for the impact of data size and time slot on the transaction throughput of blockchain.

The movie recommendation system [158] combines DP with reinforcement learning to protect user data privacy without knowing the accurate inference model. More specifically, this scheme applies DDQN that has two neural networks with the same architecture each of which consists of four FC layers to optimize the privacy budget. The state includes the item sensitivity level and similarities, and the privacy protection level. This scheme aims to increase the user privacy protection level and satisfy the QoS requirement. In the simulations based

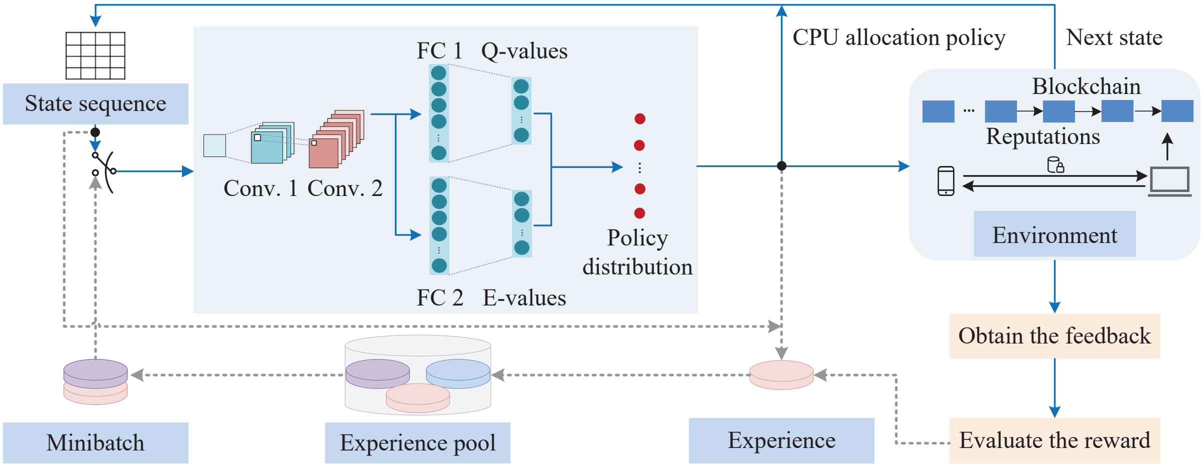


Fig. 9. RL based blockchain-assisted MEC system.

on a public movie recommendation dataset MovieLens 1M that stores 3952 movies each with 19 categories, this scheme reduces more than 28.8% privacy loss and increases more than 20.3% compared with the benchmark perturbation mechanism as proposed in [173] with the 0.2 ∼ 0.7 attack probability. The learning efficiency and user data privacy level of this scheme may degrade, if each user involves an inaccurate quantized action set.

The policy gradient RL algorithms such as DDPG, A2C, A3C and PPO have been applied in the data privacy-aware communication systems with continuous action set to address the quantization error issues. These algorithms directly generate the privacy policy instead of the policy distribution from the CNN function approximators in DQN. For instance, in the privacy-aware industrial IoT communication system as proposed in [159], the smart mobile terminal uses Ethereum blockchain to protect the sensing data from selfish, DoS and DDoS attacks, and applies DDPG to optimize the moving distance and direction of the mobile terminals in the individual offloading without cooperation. The state in this scheme contains the location of the point-of-interests and the obstacles, the coordinates of the mobile terminals, the current battery level of the mobile terminals and the sensing latency of the previous point-of-interests. The mobile terminal evaluates the reward that depends on the amount of the collected data, the geographical fairness among the point-of-interests and the energy consumption of the mobile terminals. Simulations based on three Ethereum nodes against an attacker that performs both DoS and DDoS attacks and sends 3306 abnormal requests in each time slot verify the performance gain of the proposed scheme over the benchmark.

The A3C based blockchain-enabled MEC system in [161] formulates the state that consists of the channel conditions, the available computing resources, the number of stakes and the trust value of relay nodes to choose the offloading policy, the power allocation, and the block size and interval for the discrete MDP with the reward that depends on the MEC computation rate and the blockchain transaction throughput. The A3C consists of six worker agents and a global network implemented in a central parameter server, in which a global network is used to synchronously update the network weights of the six worker agents, and each worker agent outputs the policy function via a softmax layer and estimates the value function via a linear layer. This scheme increases 50.0% computation rate and 5.9% transaction throughput compared with the fixed block interval scheme in the simulations based on a block-enable MEC system for 30 mobile devices with up to 1 W transmit power and 1 GHz CPU-cycle frequency, five relay nodes and six worker agents with 200 average transaction size and 8 MB block size. This scheme aims to protect the user data from the adversary, while it ignores the data privacy from the MEC server to the mobile devices.

Deep reinforcement learning such as PPO and DDPG helps improve the data utility of the DP-based offloading scheme in [157] that applies Laplace noise at the edge server to protect sensitive data. More specifically, PPO can enable the edge server to optimize the Laplace noise based on the state, including the data size, the sensitive level, the privacy level and the QoS to improve the long-term expected reward, including the accuracy, and the data utility and privacy. The MEC system does not consider the differences of the privacy protection level among different users, as each user may pay attention to different private contents, and has various privacy requirements for its own data at different time slots.

In another example of the RL based data privacy-aware communication similar to the MCS system in [172], multi-agent RL such as DIAL can be applied to enable each user to share their observations such as the data type and the historical data accuracy for faster optimization to increase the reward that relies on the privacy protection level, the data confidentiality and integrity, the transmission latency, and the computational overhead. As this potential scheme requires each user to share their observations, it suffers from a high communication and computational overhead and the curse of dimensionality under the high-dimensional state space and action set.

C. MmWave

The user identity and behavior privacy protection in mmWave systems can be improved by RL without relying on any training data. For example, the authentication threshold in the hypothesis test of the deep learning based authentication in [164] can be chosen based on DDQN to protect the behavior information of the radio frequency signals with reduced computational complexity. The state consists of the RSSI, channel state information, information-related amplitude and propagation delay for both the identity and behavior, and previous privacy level. Based on the current state as well as the state sequence consisting of the historical authentication thresholds and reward as the input, the online network with a hidden layer outputs the Q-values of all the feasible threshold levels. The target network designed with the same architecture updates the weights of the online network to avoid overestimation. The reward function can be formulated as the weighted sum of the recognition accuracy, the authentication accuracy, and the privacy loss.

To improve the privacy protection level of the multi-user mmWave systems, the multi-user authentication system in [165] uses a clustering method to resist spoofing attacks by capturing the self-driving heartbeat motions of users. However, the performance degrades in more complicated systems with a large number of multi-antenna users due to the higher latency in the clustering process. The PPO consisting of an actor network and a critic network can be applied to help optimize the matching parameters based on the state that can be the number of users, intermediate frequency signals, range resolution among users, and angle of arrival. The reward is the sum of the classification accuracy, false positive rate, authentication accuracy, and privacy protection level. The actor network with three FC layers outputs the mean vector of all the available matching parameters, which is used to formulate a multivariate Gaussian policy distribution for selecting the matching parameters. The critic network outputs the state value to evaluate the chosen matching parameters according to the Kullback-Leibler divergence technique.

D. Massive MIMO

Massive MIMO systems with a huge number of transmit antennas support multiple users located in different areas, and must protect the user data privacy from attackers with the QoS guarantee. For instance, the multi-cell massive MIMO communication system in [166] uses the time-fraction-wise beamforming technique to satisfy the user QoS and data privacy protection requirements based on an accurate attack model and prior knowledge of channel state information. Deep RL such as DDPG can help optimize the beamforming policy without knowing the prior information as in [166] to further improve the privacy protection performance of the dynamic systems. The state including the RSSI, inter-user interference power level, previous data protection level, and locations of all the users is applied to the actor network of DDPG. The chosen beamforming policy, the current state and the reward are saved into the experience memory pool. The BS applies the experience replay technique to randomly choose a piece of experiences from the experience memory pool, and uses these experiences as the input of the critic network that outputs the Q-values based on the energy efficiency and the throughput to evaluate the chosen policy.

Federated learning can further improve the data protection performance of the deep RL based time-division-duplex massive MIMO communication system in [166]. For example, the federated learning based compressive sensing scheme in [167] can combine DQN to choose the transmission strategy such as the amount of the uploaded learning parameters, without knowing the change model of channel impulse response from the K devices to the server and the wiretap model. The state based on the estimated legitimate channel impulse response, local training data size, and previous data eavesdropping rate, model accuracy and latency is input to the designed DQN with four FC layers, where the second and third layers are the hidden layers. The transmission strategy is chosen based on the Q-values of the feasible transmission policies under the current state according to the ϵ-greedy algorithm. According to the chosen policy, each device transmits its parameters to the server, and the server applies the stochastic gradient descent algorithm to train the global model, and sends the corresponding parameters to the participating devices. After receiving the feedback from the devices, the server obtains the immediate reward including the model classification accuracy, data privacy level of the devices, and transmission latency.

In summary, the data privacy-aware communications with the value-based RL have to quantize the privacy policy and thus may achieve a local optimum policy, due to the policy quantization errors. On the other hand, the policy gradient RL based data privacy-aware communication schemes achieve better privacy protection performance in the continuous action set, but suffer from low data efficiency and poor robustness in the learning process.

VIII. RL BASED LOCATION PRIVACY PROTECTION

Existing location privacy protection that depends on anonymization [182], blockchain-assisted consensus approach [183], information-theoretic approach [184], [185] and DP technique [186] suffers from performance degradation in a dynamic 6G system with unknown attacker mobility model. For example, the DP-based IoT location protection scheme in [186] uses Laplace noise to perturb the user sensitive location. However, the privacy budget in this scheme is chosen based on the frequent pattern records of all the users in the system, which are hard to be obtained by the IoT device. Therefore, DQN and DDPG, as important RL algorithms, have been used to improve the location privacy protection level for MEC, IRS, mmWave and massive MIMO systems against inference attacks, eavesdropping, selfish attacks and backdoor attacks, without knowledge of the attacker locations and attack patterns [176], [177], as summarized in Table VII.

A. MEC

The mobile devices with sufficient computational resources can apply DQN and deep PDS-learning to optimize the edge selection for location privacy protection under a discrete action

TABLE VII
RL BASED LOCATION PRIVACY PROTECTION

State	Policy	Reward	RL algorithms	Attacks
Blockchain data size Channel states Hash power Computing payment	Offloading policy	User data rate Location protection level Edge devices payment Computational latency Energy consumption	DQN [174]	Inference attacks
Battery level Energy consumption	Edge device Offloading policy	Task dropping probability Privacy loss rate	PDS [175]	Eavesdropping
Vehicle driving direction Required caching resources	Caching policy Delivery policy	Location protection level Successful caching rate Energy consumption Content delivery latency	DDPG [176]	Selfish attacks
Vehicle location Sensitivity level Attack strength history	Privacy budget Perturbation angle	Privacy level QoS loss	DDPG [177]	Inference attacks
Perturbed channel states Energy demand vector	Transfer power vector IRS phase shifts	QoS Location protection level	DDPG [178]	Inference attacks
Current human activities Sampling data size Task type Feature dimensions	Classification threshold	Recognition accuracy Tracking accuracy Latency Location protection level	Dyna-Q [179]	Inference attacks
Number of users Attack probability Current datasets Data confidence	Beamforming policy	Data confidence Location protection level Benign rate Accuracy	DQN [180]	Backdoor attacks
User-BS channel states Number of users Corresponding antennas Attack probability	Privacy budget	Mean squared error Symbol error rate Execution time Location protection level	DDPG [181]	Inference attacks

set. For example, the RL based privacy-aware MEC communication [174] applies DQN with two hidden layers to determine whether to offload tasks to the available edge devices against the attacker that can predict the newly generated task size to infer the MEC server location and usage patterns. This scheme quantizes the action set as the number of feasible offloading decisions for each IoT device, which offloads the task to the edge device if the offloading decision is one, and processes the task locally otherwise. In this scheme, the state consists of the blockchain data size, the channel states among the mobile device and edge devices, the hash power and the computing payment. The reward depends on the user data rate, the user location protection level, the payment received by the edge devices, the computational latency, and the energy consumption of the mobile device. In the simulations, the Biokin sensors are used as IoT devices to collect health data such as motion data 50 ~ 150 KB every second and send these data to the 10 mobile devices as miners for medical services. The mobile devices with 10 KB block size offload data to the MEC server with a computational capacity of 1 0 GHz/sec and 50 mW static circuit power via Wi-Fi wireless communication following IEEE 802.11g. The results show that the proposed scheme improves the privacy level by 12.7% as compared with the benchmark CMDP scheme, but suffers from a low learning accuracy under the time-varying and continuous action set.

Another location privacy-aware IoT offloading scheme named PAO as proposed in [175] applies deep PDS-learning to choose the edge devices and offloading policy from the action set that includes all the available edge devices and the quantized offloading policy levels. The DQN that consists of six FC layers uses the state as the input, including the battery level and the offloading energy consumption of the IoT device. This scheme assumes that the channel model from a mobile device to an edge MEC server follows a log-normal distribution. For the IoT device that generates 1000 bits of data at 1/700 of the computing speed of the three MEC servers with 1 MHz uplink bandwidth, PAO saves the convergence time by 86.7%, reduces the task dropping probability by 25.0% and decreases the privacy loss rate by 15.2% as compared with the DQN based offloading scheme, after 200 time slots from the beginning of the game against the attacker.

DDPG helps the vehicle and the mobile device optimize the privacy budget, the perturbation angle, and the IRS phase shifts without quantizing the action set and thus further improve the location privacy protection level compared with DQN and deep PDS-learning. For instance, a vehicular edge computing system in [176] applies DDPG to choose the edge caching and content delivery policy. The reward function is formulated with

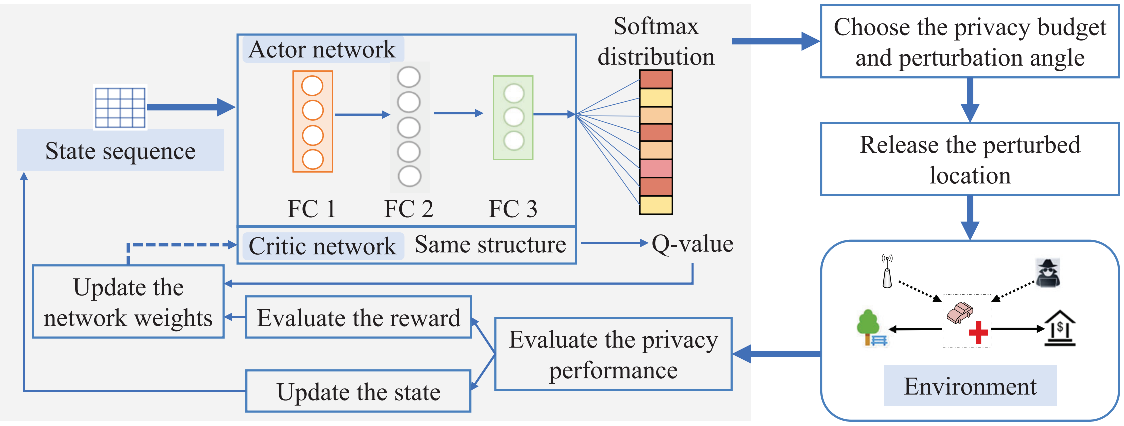


Fig. 10. Deep RL based location privacy protection.

the vehicle location protection level, the successful caching rate, the energy consumption and the content delivery latency. The mobile device formulates the state based on the vehicle driving direction and the required caching resources. This scheme designs four DNNs each of which has three FC layers, in which the primary actor network directly outputs the edge caching and content delivery policy, the primary critic network evaluates the chosen policy and outputs the evaluated Q-values, and both the target actor and critic networks generate target Q-values to train the two primary networks. The system reshapes the discrete edge caching policy (i.e., zero or one) into a continuous action set, and uses the max-weighted bipartite matching obtained by the Hungarian algorithm to estimate the edge caching policy. In the simulations based on a real-world dataset from Uber with 100 vehicles each of which has 5 GB caching capacity and 10 MHz bandwidth to cache the content data with size ranging from 10 to 50 MB, the proposed scheme achieves 86% successful content caching probability, which is 10.3% higher than the greedy content caching scheme.

The deep RL based location privacy protection scheme as proposed in [177] further improves the privacy protection level for sensitive location-based service systems by incorporating DP and DDPG to optimize the privacy policy that consists of the privacy budget and the perturbation angle against the adversary that can infer the user preferences and life patterns. As illustrated in Fig. 10, the state that contains the vehicle location, the sensitivity level of the current location and the attack strength history estimated by the vehicle is used as the input of the actor network with three FC layers that outputs the privacy policy. According to the experience replay technique, a critic network with three FC layers outputs the Q-value of the chosen policy to update the weights of the actor network, and thus evaluate the effectiveness of the chosen privacy budget and perturbation angle. After obtaining the policy, the vehicle adds gamma distributed noise to the original location, based on the privacy budget and perturbation angle to obtain the perturbed location. In the simulations based on a 6×6 km^2 square map with four sensitive locations, this scheme improves the privacy level by 77.5% and reduces the QoS loss by 7.7%, as compared with the Geo-indistinguishability scheme after 1500 time slots, given the semantic model and time-varying sensitivity model of each location.

B. IRS

IRS helps improve the data transmission quality, but has location leakage vulnerability to the adversary that can intercept the legitimate data as well as infer the locations of both the IRS and users. This issue can be mitigated by combing DP with deep RL to choose a location protection policy such as the IRS phase shifts, without knowing the accurate adversary model and channel changing model. For example, the RL based DP-aided location protection scheme in [178] applies the DDPG to jointly optimize the transfer power vector of the energy transmitters and the IRS phase shifts based on the state that includes the perturbed channel coefficients among the energy transmitters, users, IRS and IoT devices and the energy demand vector of the IoT devices to increase the reward, including the QoS and the location protection level of the energy transmitters. By assuming that the harvested energy is enough to support the data transmission, this scheme uses FC layers to replace the Conv. layers in the traditional DDPG for less space complexity. Simulations are performed with $3 \sim 17$ static energy transmitters, each of which has power ranging from 20 to 50 dBm and an IRS having 20 available reflection phase shifts, showing that the proposed scheme improves the reward by 29.4% compared with the random power allocation scheme with 10 IoT devices.

C. MmWave

The multiple-inhabitant activities will cause serious user location leakage in mmWave systems. Thus, a 79 GHz mmWave recognition system with six users in [179] applies deep learning to obtain multiple user tracks and human activities. It has severe location leakage without enough labeled datasets. In this case, the system can apply Dyna-Q to choose the classification threshold from the action set quantized from zero to one, based on the state including the current human activities, sampling data size, task type, feature dimensions, and previous location leakage level. The goal is to maximize the reward that depends on the activity recognition accuracy, user tracking accuracy, latency, and user location protection level. A number of virtual location protection experiences including the simulated state-action pairs, corresponding simulated reward, and next state, are generated by a Dyna architecture and used to update the Q-values with real experiences, thus avoiding random exploration in the initial learning process. However, this system may fail in a mmWave system with hundreds of users without quantizing the available threshold levels. This issue can be addressed by deep RL such as DDPG, which uses neural networks (e.g., CNN) to extract the protection features and further improve the user location protection performance, thus reducing the quantization errors in the continuous action set.

Another user location protection mmWave system in [180] combines federated learning with deep learning to choose a beamforming policy, using the dynamic norm clipping method based on the accurate backdoor attack model to protect user location privacy. The performance can be further improved by deep RL, without knowing the attack interval and strategies. More specifically, the server applies DQN with two Conv. layers and two FC layers to optimize the beamforming policy, in which the state (consisting of the number of users, probability of attacker existence, current datasets of all the users, probability of the location leakage, and previous data confidence) is used as the input. By using the neural network function approximators, the DQN outputs the Q-values of all the feasible beamforming policies under the current state, which relies on the reward function formulated with the data confidence, user location protection level, benign rate and accuracy.

D. Massive MIMO

Due to a large number of users and their various locations, massive MIMO systems are vulnerable to inference attacks. A location privacy-preserving channel estimation algorithm designed in [181] incorporates DP and the PHY-layer signal processing techniques such as the channel estimation to protect the user location information. The Frank-Wolfe and singular value decomposition is applied to estimate the user-BS channel states, without revealing any location information of the users. Nevertheless, this algorithm relies on the accurate channel model and the adversary strategies, thus cannot satisfy the DP requirement under the massive MIMO based 6G system with large-scale dynamic users. Thus, the BS can apply an RL algorithm such as DDPG to optimize the privacy budget in the DP, based on the state including the estimated legitimate user-BS channel states, number of users and corresponding antennas, estimated attack probability, and previous location protection performance. This system aims to maximize the reward that relies on the normalized mean squared errors of the channel estimation, symbol error rate, execution time, and average location protection level of all users.

In summary, the existing RL based location privacy protection schemes are designed for the MEC and IRS systems against inference attacks, eavesdropping, and selfish attacks. The location protection performance of mmWave and massive MIMO systems can be improved by Dyna-Q, DQN and DDPG, which enable wireless devices to optimize the classification threshold, beamforming policy, and privacy budget, without relying on an accurate attack model.

IX. RL BASED UAV SECURITY AND PRIVACY PROTECTION

RL based UAV security and privacy protection schemes address jamming, eavesdropping, differential attacks and DDoS attacks and apply RL to optimize the transmit power and video layer selection, the payment policy and the trajectory planning to improve the security performance and the privacy protection level [196], [197]. As summarized in Table VIII, these schemes rely on the performance history and the current channel states to increase the long-term expected reward consisting of the peak signal-to-noise ratio (PSNR), energy efficiency, and packet arrival rate. RL algorithms such as WoLF-PHC and DDPG enable UAV to enhance security under dynamic network environments without relying on the accurate attack pattern information.

A. Anti-Jamming UAV Communications

Due to the limited battery capacity and high mobility, UAVs have to address jamming attacks, especially smart jamming, and the performance can be improved by RL. For example, the UAV controller communication in [88] that applies Q-learning and WoLF-PHC to choose the UAV transmit power vector over the available frequency channels based on the previous attack mode increases the SINR, the safe rate and the secrecy capacity with less energy consumption against jamming. The UAV communication performance under a large transmit power range and available frequency channels can be improved by the power allocation based on DQN aided by the stochastic gradient descent algorithm. This scheme depends on the CNN architecture that consists of two Conv. layers and two FC layers: Conv. 1 has 20 filters each with size 3×3 based on the length of the reshaped state sequence, Conv. 2 involves 40 filters each having a size 2×2, FC 1 has 180 neurons that equal to the output of Conv. 2, and FC 2 with size 64 outputs the Q-values of the 64 feasible policies. For the UAV transmission with up to 400 mW transmit power against a smart jammer with up to 400 mW jamming power, this scheme increases 11% safe rate from the Q-learning based scheme, but the UAV transmission may have a high outage probability, thus cannot satisfy the QoS requirements in the multi-UAV system with multiple smart jammers.

Therefore, a knowledge RL based anti-jamming UAV communication scheme is proposed for the large-scale swarm communications indicating a large state space in [187], which enables the UAVs to exploit the anti-jamming features in the state with the domain knowledge technique and thus accelerate the convergence speed. More specifically, a UAV system applies DDPG to jointly select the channel power allocation policy and the trajectory of all the UAVs based on the state that is formed with the position information, the maneuvering state of UAVs and the channel state. The

TABLE VIII
RL BASED UAV SECURITY AND PRIVACY PROTECTION

State	Policy	Reward	RL algorithms	Attacks
Previous attack mode	Power allocation	SINR Safe rate Secrecy capacity Energy consumption	Q-learning WoLF-PHC DQN [88]	Jamming Eavesdopping Spoofing
UAV position Channel states Maneuvering state	Power allocation Trajectory control	SINR QoS requirement Energy consumption	DDPG [187]	Jamming
Previous PSNR Task priority Channel states Jamming power	Coding rate Modulation type Power allocation Quantization parameter	PSNR Throughput Transmission latency Energy consumption	Q-learning Safe DQN [188]	
Robot location Previous BER RSSI Energy consumption	Relay policy Moving distance	SINR Energy consumption	Q-learning Safe DDQN [189]	
UAV position Available users	AN strategy Power allocation Moving velocity	Secrecy rate Energy consumption Map limitation penalty	MADDPG [190]	Eavesdropping
UAV position	Power allocation Trajectory control Video layer selection	PSNR Energy consumption	Safe DQN [191]	
Channel states Jamming power Previous latency	Power allocation Network coding policy	Eavesdropping rate Transmission latency BS energy consumption	Dyna-Q [68]	
Model update quality	Payment level Local update strategy	Sensing cost Aggregate accuracy Energy consumption Model update quality	Q-learning [192]	Differential attacks
UAV behaviors	Detection policy	Detection accuracy	DDPG [193]	Intrusion attacks
User offloading policy User transmit power Energy consumption	Power allocation UAV selection	Penalty counter Energy consumption	DQN [194]	Inference attacks
UAV position Remaining data payload	Device selection Power allocation Deployment strategy Subchannel selection	Execution time Learning accuracy loss	A3C [195]	

reward function is formulated with the SINR, the moving energy consumption, and the QoS requirement. A UAV system without the physical collisions among UAVs is simulated in Python 3.6 software equipped with pytorch1.4 with 5 UAVs each of which sends messages with 400 mW transmit power and flies at 500 m height within a 4 km × 3 km area against a reactive jammer with three available jamming channels and maximum power 400 mW. Simulation results show that this scheme saves the average energy by 18.3% as compared with the SARSA-Q based anti-jamming scheme with both the UAVs and the jammer having constant flying velocities.

Compression and coding parameters such as the video quantization parameter and the channel coding rate are highly related to the UAV video transmission performance. Nevertheless, existing video transmission schemes mainly focus on fixed compression and coding parameters and thus fail to meet the various video quality-of-experience requirements with dynamic UAV channel states under smart jamming. Thus, the RL based video transmission scheme in [188] formulates the state with the task priority, the UAV-ground channel state, the received jamming power at the ground station, the PSNR of the UAV video signals, the transmission latency and energy consumption, and the throughput against smart jamming. The UAV applies fast Q-learning to optimize the video quantization parameter, the channel code rate, the modulation type and the transmit power based on the state, which aims to maximize the long-term expected reward that relies on the PSNR, and the transmission latency and energy consumption.

To further improve the learning efficiency for the UAV with enough energy resources, this video transmission system applies safe DQN to guarantee the video quality-of-experience, which uses the SINR as the security criterion to avoid ex-

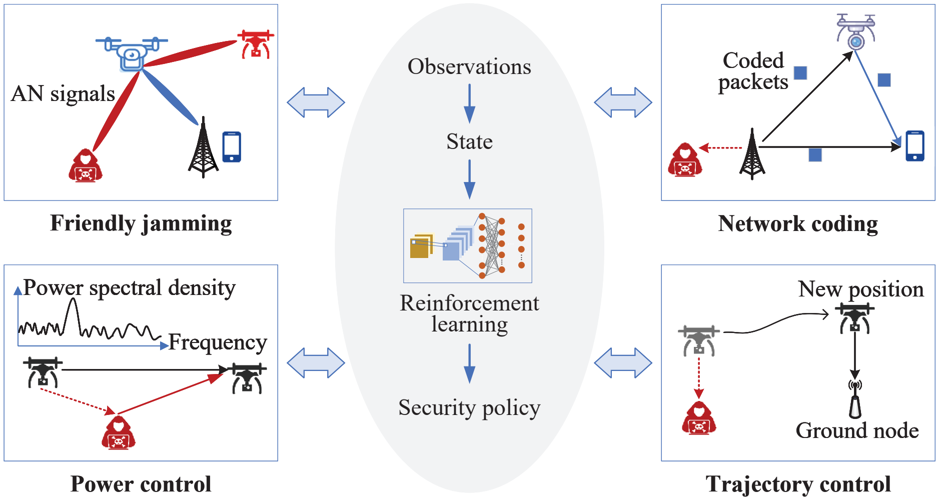


Fig. 11. RL based UAV secure communications against eavesdropping.

ploring the risky transmission policies that cannot satisfy the given threshold. Two CNNs (i.e., the Q-network that outputs the Q-values and the E-network that evaluates the E-values) sharing the Conv. layers are designed to formulate the policy distribution based on the modified Boltzmann distribution. In the simulations, the UAV at a 10 m/s speed compresses the captured video with the compression coding standard H. 264 and sends the video stream with binary phase-shift keying, quadrature phase-shift keying or 16-quadrature amplitude modulation and power ranging from 100 to 180 mW. Each of the two CNNs has 20 and 40 filters in two Conv. layers, respectively, to guarantee the video transmission quality. The proposed scheme improves the SINR by 28.6%, reduces the latency by 47.6%, and saves the energy consumption by 62.2% compared with the adaptive modulation scheme against a random jammer with feasible power changing from 100 to 120 mW. Nevertheless, this scheme ignores the impact of video stream data size and UAV mobility on the video transmission performance, which may result in transmission failure under a highly dynamic UAV system.

Robots can help further improve the UAV-ground communication performance of the RL based anti-jamming schemes. For example, the robot relay scheme proposed in [189] combines tile coding technique (i.e., a function approximation approach) with Q-learning to optimize the robot relay policy based on the state that contains the robot location and transmission energy consumption and the previous transmission quality. Compared with typical Q-learning, this scheme maps the state into a number of anti-jamming features for less storage overhead and aims to increase the SINR and save the robot energy. The robot with sufficient resources can also apply DDQN equipped with a risk network, an online network and a target network. Each of the three networks includes an input layer with 5 neurons, a hidden layer with 256 neurons and an output layer with 27 neurons that relies on the number of feasible relay policies. A Boltzmann policy distribution is formulated with the Q-values and the E-values that rely on the BER of the messages and the QoS requirement. Simulations are performed with a UAV sending messages with 100 mW power to a BS and a robot relay with transmit power changing from zero to 200 mW and moving within 5 m to resist a smart jammer with maximum power 10 mW. The results show that the proposed scheme achieves 36.6% less outage probability, 67.5% lower BER, and 31.4% less robot energy consumption than the Q-learning based trajectory control scheme. However, this scheme requires the robot with enough resources to support DNNs, and suffers from anti-jamming communication performance degradation if the robot mobility is slower than the UAV.

B. *Secure UAV Communications*

The line-of-sight of UAV channel links improves UAV communication performance but also increases the intercept probability of the eavesdroppers, especially the active eavesdroppers that can send jamming signals based on the observed channel states and thus induce the UAVs to increase their transmit power. Therefore, the RL based secure communication can be used to decrease the intercept performance of the eavesdroppers compared with the traditional convex optimization schemes in the optimization of the friendly jamming, moving strategy and attack detection mode in UAV systems, as illustrated in Fig. 11. For example, the UAV can apply RL such as A2C to optimize the secure communication policy and use the network coding technique (such as the random linear network coding) to encode each message into several independent encoded packets, thus hiding messages in multiple data flows against eavesdropping [198].

A multi-agent RL based cooperative secure UAV communication scheme is proposed in [190], which enables the UAV network to apply MADDPG to choose the secure communication policy, including the transmit power, friendly jamming power and the moving velocity of each UAV. With the goal of maximizing the long-term expected reward that relies on the secrecy rate, the map limitation penalty, and the energy consumption of all UAVs, this system formulates the state with the UAV position and the index of the objective ground users. In this system, each UAV has an actor network and shares a global critic network. More specifically, the state is input to the actor network of each UAV that directly outputs the UAV secure communication policy. With the state and the chosen UAV secure communication policies of all the UAVs, a global critic network with five FC layers evaluates the corresponding Q-values and updates the weights of each actor network. Different from the traditional MADDPG, this scheme designs an attention network to extract the secure communication features with higher importance to the system, which is added to connect with the global critic network to reduce the exploration space. In the simulations based on a 100 m × 100 m square area, a UAV as the transmitter and three UAVs as friendly jammers flies at height from 15 to 50 m and sends messages to six ground users with up to 2.1 W transmit power against three eavesdroppers. The results show that this scheme improves 11.1% secure rate compared with the DDPG based secure communication scheme. Nevertheless, this scheme ignores the observation sharing among the UAVs and has learning efficiency degradation in large-scale and dynamic networks.

UAVs with insufficient computational and energy resources have to avoid communication failure and guarantee the QoS for the live streaming applications under eavesdropping. Safe RL such as safe DQN can reduce the dangerous exploration in the optimization of the video secure communication policy such as the video layer and the transmit power. Therefore, the UAV-enabled video streaming transmission scheme in [191] with each UAV flying at a fixed altitude applies safe DQN to select the video layer that represents the requirement of the PSNR, the structural similarity and the data rate, the transmit power and the trajectory based on the UAV locations (i.e., the state). In this scheme, the safe DQN has a Q-network and two target Q-networks, in which the Q-network with the state as the input that has four FC layers estimates the long-term expected reward and the two target Q-networks are used to evaluate the auxiliary constraint cost. This scheme formulates the secure communication process as a CMDP and uses Lyapunov theory to build a safe policy set. The reward depends on the PSNR, and the UAV transmission and moving energy consumption. Simulations are performed in a Python platform, in which 2 ∼ 6 UAV jammers send AN signals to the ground with 0 ∼ 10 W power and a UAV at 100 m height sends video streaming with up to 200 mW transmit power, 100 KHz bandwidth and five video layers to the ground. The results show that this scheme improves the energy efficiency by 12.5% compared with the Lagrangian-based DQN algorithm under 4×10^{-4} learning rate.

Network coding provides inherent data protection for UAVs by mixing the information from different data flows. For example, the RL based UAV-aided secure communication scheme in [68] applies Dyna-Q to optimize the number of coded packets, the packet allocation and the transmit power based on the state including the legitimate channel gain, the jamming power received at the UAVs, and the transmission latency. This scheme designs a Dyna architecture that consists of a virtual model generating a number of simulated experiences, which uses to update the long-term expected reward that includes the eavesdropping rate, the transmission latency and the energy consumption. The performance gain is verified in a simulated secure communication scenario, where a BS sends 7000 bytes of picture data coded by the random linear network coding algorithm with power changing from zero to 400 mW to the mobile device, and three UAVs help relay the coded packets with fixed power 20 mW against an active eavesdropper that uses intelligent devices to wiretap the channel state and applies Q-learning to optimize the jamming power from zero to 400 mW. For instance, the proposed scheme achieves 87.2% lower intercept probability and 84.9% lower outage probability after 2500 time slots than the conventional relay selection scheme. This scheme quantizes the secure communication policies into a discrete action set and relies on accurate reward signals and state observation. It can suffer from slow learning speed and quantization errors under dynamic UAV networks with the delayed feedback over limited bandwidth.

C. *Privacy-Aware UAV Communications*

RL can be combined with several new techniques such as federated learning, blockchain and DP to protect the UAV private and sensitive information from malicious attacks during the offloading and crowdsensing processes. For example, the UAV-assisted MCS privacy-aware communication system proposed in [192] combines federated learning with Q-learning to select the payment level of the task publisher and the local model quality strategies of UAVs to maximize the reward including the sensing cost, the energy consumption of each UAV, the quality of local model update of the participate UAVs, and the accuracy of the aggregated global model. The state consists of the previous quality of local model update sequences of participating UAVs. This system uses blockchain to record the UAV behaviors with a PoW based incentive mechanism against differential attacks, low-quality local model update attacks and contribution records tampering attacks, and applies the local DP mechanism and federated learning based on the stochastic gradient descent algorithm to protect UAV data privacy following the aggregate accuracy constraints. Simulations are performed in the MCS system with 4 BSs and 80 ∼ 120 UAVs flying at 100 ∼ 300 m height in a 1000 × 1000 m^2 terrain area. The results show that the scheme improves the average quality of the local model update by 40.0% compared with the randomized policy scheme. However, the optimization time in this scheme is significantly raised due to the increment of the UAV number and the task number, which sometimes can be even longer than the task latency requirement.

To reduce the quantization errors in the UAV edge computing network against smart attacks, the RL based intrusion detection scheme in [193] applies DDPG with two primary networks and two target networks to optimize the malicious attack detection policies of all the UAVs to maximize the long-term expected reward that depends on the detection accuracy. More specifically, the primary actor network uses the state including the behaviors of UAVs as the input and directly outputs the detection policy of each UAV, and the primary critic network evaluates the chosen policy and updates the weights of the primary actor network. Two target networks

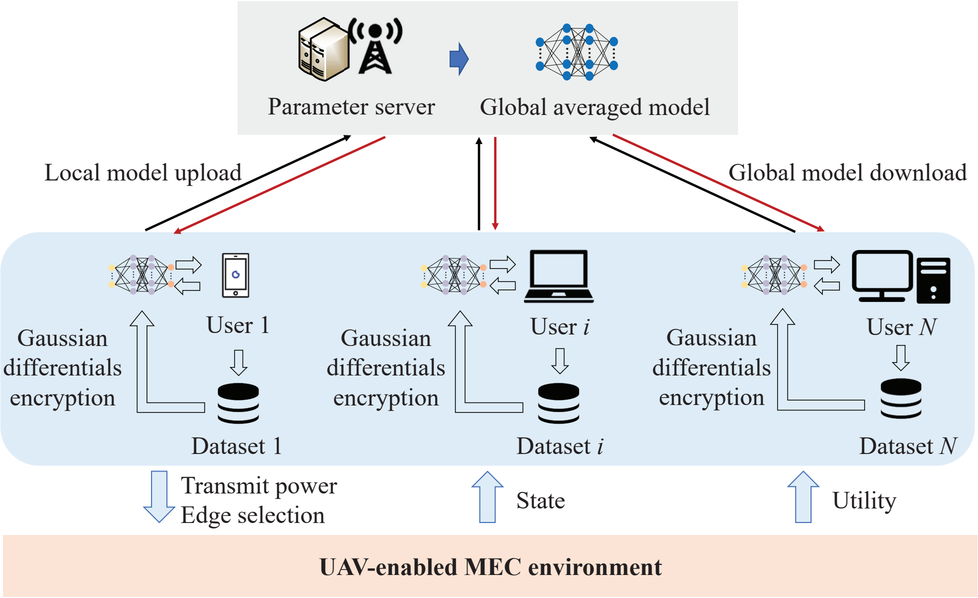


Fig. 12. RL based UAV-enabled MEC data privacy protection with federated learning.

are used to update the weights of the two primary networks to avoid choosing the local optimum policies. Simulation results based on a UAV network within a $500 \times 500 \times 500$ m^3 area with 5 smart attackers, 50 mobile users, and 20 UAVs with transmit power up to 40 dBm show that the proposed scheme achieves 95.0% detection accuracy, which is 48.4% higher than the opportunistic scheme after 1000 time slots. The system does not account for the detection overhead (such as the latency and the UAV moving energy), thus may fail to detect the smart attacks in practical UAV networks without enough energy resources.

A multi-agent federated RL based semi-distributed resource management scheme is proposed in [194] to combat server attacks and massive privacy leakage. As illustrated in Fig. 12, this scheme combines multi-agent DQN and federated learning algorithm to choose the number of UAVs used to offload tasks and the corresponding transmit power of mobile users to increase the reward consisting of the penalty counter and the energy consumption of both the mobile users and selected UAVs with sensitive data protection guarantee. The state includes the task offloading policy of each UAV, the current transmit power of each mobile user and the system energy consumption, which is used as the input of the DQN with three hidden layers. In this scheme, each mobile device does not share the learning experiences with others, and may spend a lot of time randomly exploring unnecessary policies at the beginning of the learning process. This scheme has two parts: The single-agent training process applies standard DQN to select the offloading policy, while the Gaussian differentials encryption uses DP to encrypt the experiences in the dataset as well as the network parameters in the user local model. By applying the experience replay technique, each UAV randomly chooses 256 experiences from the replay memory pool with a size of 2000 for higher privacy protection performance. Simulation results with 20 $\sim$ 200 mobile users with transmit power up to 20 dBm and 10 UAVs flying at 50 m height and 4 m/s speed show that the proposed scheme can reduce the total UAV energy by 38.2% compared with the benchmark without resource management scheme with 200 mobile users.

The deep RL based privacy-aware UAV resource allocation system as proposed in [195] combines A3C with federated learning to jointly optimize the device selection (e.g., the mobile device, the vehicle and the IoT device) and the deployment, the subchannel selection and the power allocation of each UAV, which enables the devices to process the raw sensitive data locally rather than offload it to the UAV edge servers for user privacy protection. More specifically, each UAV formulates the state based on its horizontal location, the location of the selected devices and the remaining data payload, which is input to both the actor and critic networks. This scheme aims to increase the reward including the learning accuracy loss and the execution time. Instead of updating the weights with the target networks, this system uses federated learning to train a global model to update the two networks of each UAV for higher learning efficiency. Each network has three FC layers that involve 256, 256, and 128 neurons, respectively. In the simulations with four UAVs at 150 m height with power up to 150 mW and 100 devices with 50 mW transmit power, this scheme achieves about 91.0% learning accuracy, which is 7.1% higher than the benchmark without device selection. However, the system directly applies the standard A2C that relies on immediate experiences without considering the policy selection priority, and thus explores local optimum privacy protection policy due to any inaccurate weights update of the global network.

In summary, the value-based RL including Q-learning and DQN, the policy gradient RL in terms of DDPG and A3C, and the multi-agent RL such as WoLF-PHC have been applied in the selection of UAV security policies. Nevertheless, these schemes rely on either the complete state observation and

TABLE IX
RL BASED CROSS-LAYER SECURITY AND PRIVACY PROTECTION

Learning agent	Policy	Reward	RL algorithms	Layers
SDN controller	Link selection [199]	Reverse delivery ratio Trust value of each vehicle	DQN	PHY Network
Edge device	Bandwidth allocation [200]	Movement authority Boundary probability Bandwidth of each train	A3C	PHY
Coordinator	Encryption key size Power allocation IRS phase shifts [79]	SINR Eavesdropping rate Transmission latency Data protection level Sensor energy consumption	Safe Dyna-Q PPO	PHY MAC
SDN controller	Computational resource allocation [201]	Privacy loss QoS of users	SARSA-Q	Network Application
Edge device	Vasopressor dosage Intravenous policy [202]	Sequential organ failure assessment score	DDQN	PHY Network
User	Bitrate of video chunks [203]	Quality of experience	A3C	PHY Application
MEC server	Privacy protection algorithm Block interval Data flow selection [204]	Energy consumption Transactional throughput	A3C	Network
	Protection level User location [205]	Charging time Offloading time	DQN	
	Caching policy [206]	Cache hit rate	DDPG	

immediate reward signals from the environment or the fixed UAV location and trajectory, which may have difficult implementation in practical UAV networks under more intelligent attacks.

X. RL BASED CROSS-LAYER SECURITY AND PRIVACY PROTECTION

The wireless security solutions at PHY-layer, MAC layer and network layer can cooperate to improve the security against smart attacks that change the attack modes and location [22], [207]. The corresponding cross-layer security policies (e.g., the encryption key in the MAC layer and the transmit power in the PHY-layer) depend on the accurate attack mode and the security strategies in each layer, which are rarely known by the mobile devices and BSs [79]. Therefore, RL such as Dyna-Q, SARSA-Q, DQN and DDPG can be applied to optimize the cross-layer security policies under the time-varying attack model and channel states. Existing RL based cross-layer security and privacy protection schemes are summarized in Table IX.

A. *Cross-Layer Security*

Eavesdroppers and man-in-the-middle attackers can simultaneously attack the multiple layers including the PHY-layer, MAC layer and network layer in 6G cellular systems, which makes the existing RL based secure communication at the PHY-layer or the network layer suffer from performance degradation [208]. Therefore, a software-defined trust based vehicular architecture in [199] is designed to collect the physical link and network link information to resist malicious attacks (e.g., man-in-the-middle attacks) at both the PHY-layer and the network layer, which includes a trust information module, a storage module, a transaction management module, and a learning module. More specifically, the learning module in the software-defined networking (SDN) controller uses DQN consisting of an input layer, two Conv. layers, two hidden layers and an output layer to optimize the link selection policy of the source vehicle for secure communication based on the state that contains the reverse delivery ratio and the trust value of each vehicle. In the DQN structure, Conv. 1 has 32 filters with 2×2 kernel size, Conv. 2 with the same kernel size as Conv. 1 has 48 filters and the two hidden layers have 512 neurons. The system maximizes the long-term expected reward that includes the reverse delivery ratio and trust value of each vehicle to avoid communication failure among connected vehicles, by exploring all the available state-action pairs including the dangerous exploration that results in transmission failure even network disaster. Simulations are performed based on $8 \sim 32$ vehicles with $1 \sim 11$ Mbps data rate, the results show that the proposed scheme improves the reward by 41.0% with 32 vehicles compared with the CNN-based security scheme.

Due to the open protocols and network instability, the communication-based train control systems are easily threatened by network attacks such as the man-in-the-middle attackers that can steal the transmitted data or send malicious information to users. Therefore, the RL based cross-layer secure communication scheme that includes the detection and defense stages is proposed in [200] to resist smart attacks that can falsify the movement authority of the systems.In the detection stage, the system uses the movement authority as the detection basis and combines the long short-term memory (i.e., a kind of recurrent neural network) with the support vector machine model based on the labeled dataset to detect malicious users. In the cross-layer defense stage, the edge device applies A3C with 8 worker agents and a global network to choose the bandwidth allocation policy of each edge server to the total trains with the goal of maximizing the long-term expected reward that relies on the bandwidth of each train, movement authority, and the boundary probability. The state of A3C consists of the position of trains, the position of the front train, the movement authority sequence, and the confidence rate. This scheme improves the accuracy rate by 76.5% compared with the traditional intrusion detection scheme in a communication-based train control system with 8 trains each of which moves at a speed up to 80 km/h, transmits messages with 44 dBm power and owns 5 MHz bandwidth totally.

To improve the cross-layer secure communication performance against active eavesdropping in the healthcare sensing data transmission, the RL based sensor encryption and power control scheme proposed in [79] formulates the reward with the transmission latency, the sensor energy consumption, the SINR of sensor signals, the eavesdropping rate and the data protection level. More specifically, the coordinator uses the SINR of the sensor signals as the security criterion to enable the safe exploration in the selection of the encryption key size and transmit power of the sensor, and the IRS phase shifts. The transmission policy is chosen with both Dyna-Q and PPO based on the state containing the priority of the healthcare sensing data, the received jamming signal strength and the channel states of both the sensor and the IRS. In the PPO based secure communication system without considering the vulnerable exploration in the continuous action set, an actor network involving four FC layers outputs a 32-dimensional mean vector of the feasible policies to formulate a multivariate Gaussian policy distribution, and a critic network with four FC layers evaluates the one-dimensional advantage value of the chosen transmission policy, as shown in Fig. 13. In the actor and critic networks, both the two hidden layers have 64 neurons and the input layer has 65 neurons. According to the chosen policy, the sensor processes the healthcare sensing data in analog signals into digital signals based on an A/D converter, encrypts the data with the selected key size, and transmits the encrypted data with the chosen power level to the coordinator. An electroencephalography system is simulated with a sensor collecting sensing data encrypted based on the advanced encryption standard with three priorities, a coordinator with transmit power ranging from 0.1 to 1 mW and a jammer with jamming power up to 0.02 mW. The results show that the proposed scheme achieves 49.1% lower eavesdropping rate and 60.1% lower intercept probability compared with the IRS-aided secure wireless communication.

B. Cross-Layer Privacy Protection

The value-based RL (e.g., SARSA-Q and DDQN) and the policy gradient RL (such as A3C and DDPG) have been applied to optimize the cross-layer privacy protection policy such as the video chunks bitrate at the PHY-layer and the data flow selection and caching policy at the network layer against the man-in-the-middle attacks and eavesdropping.

1) Data Privacy-Aware Communications: With the limited storage and computational resources, network instability and hierarchical infrastructure, MEC systems are vulnerable to malicious attacks that can steal the sensitive data in the data transmission process among the servers and mobile devices. Existing data privacy protection schemes such as the quantum encryption based cross-layer authentication in [209] rely on the accurate attack channel states and eavesdropping patterns, which suffer from data leakage in hierarchical MEC networks with time-varying channel conditions.

This issue can be addressed by the RL based dynamic customizable privacy-preserving model as designed in [201], which applies SARSA-Q to optimize the computational resource allocation policy for the SDN controller to maximize the reward relying on the QoS of users and the privacy loss against the smart attacker that collects the private information of the mobile devices and destroys the location privacy and identity. The MEC server observes their regions, the message published by each mobile device, the time slots, and the participating users. Simulations are implemented on the Java platform and performed with the Yelp dataset consisting of more than 188,000 businesses and about 6 million user reviews. The results show that the proposed scheme reduces privacy loss by 38.0% compared with the benchmark scheme. However, this scheme quantizes the state space and defense action set, which results in quantization errors and thus slows the learning speed in the complicated MEC systems with a large number of mobile devices and servers.

With a large number of personal health information such as electronic medical records, the MEC servers in clinical decision support systems that help doctors make treatment decisions have to resist the curious and malicious nodes in both the PHY and higher layers. For example, a privacy-preserving edge-computing-enabled clinical decision system proposed in [202] combines DDQN with a federated learning algorithm to choose the vasopressor dosage and intravenous policy based on the state consisting of the electronic medical records, which are divided into 47 features including Demographics, Lab Values, Vital Signs, Intake, and Output Events, etc. By exploring all the feasible policies that contain the risky policies related to severe privacy leakage under the current state, this system aims to maximize the long-term expected reward formulated based on the sequential organ failure assessment score, representing the lactate levels of the patient and the extent of the organ failure. More specifically, the edge device applies DDQN that has two hidden layers with both 128 neurons on the fully decentralized federated framework to choose sequential clinical treatment policy and uses homomorphic encryption in the training process to further improve the secure transmission performance. Simulations are performed on the Python 3.8 platform with a 1.25 TB SSD cache and 1 ~ 8 MEC servers with 100 MB/s bandwidth, which shows that the proposed scheme provides the trust regions in real clinical decision support systems. As for implementation, this scheme has to

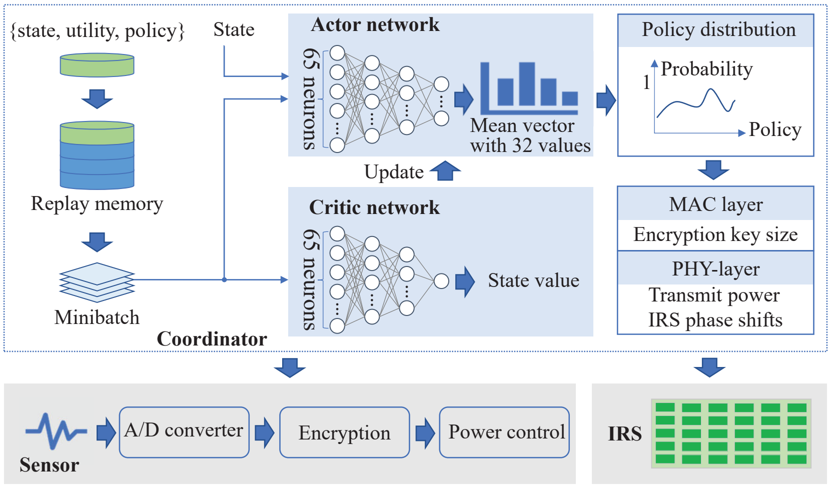


Fig. 13. RL based cross-layer secure healthcare sensing data transmission against active eavesdropping.

divide the PHY cross-layer security action set accurately and further improve the learning efficiency with a large number of patients.

A cross-layer privacy-aware communication scheme is proposed in [203] to jointly optimize the PHY-layer resource allocation and the application layer rate adaptation policy, in which the BS applies convex optimization to optimize the beamforming policy and the user uses A3C with a critic network and an actor network to select the bitrate of video chunks to increase the quality of experience of each chunk. The state is formulated with the measured network throughput, the downloading time of the previous video chunks, the complexity of the current video chunk, the sizes of video chunks, the current buffer occupancy, the number of remaining chunks in the video and the estimated previous video quality. Both the two CNNs in the A3C have a one-dimensional Conv. layer with 128 filters as the input layer, a hidden layer with 128 neurons, and an output layer with 6 neurons equalling the number of the available encoding bitrates. In the multi-cell scenario with 12 cells each having a BS with three antennas and 12 2-antenna users, the proposed scheme improves the quality of experience by 25.0% and the PSNR by 13.3% and reduces the latency by 91.3% compared with the weighted sum mean-square error minimization scheme. This scheme relies on complete state information and accurate reward signals, which may result in data leakage and video transmission failure in a partial observation scenario.

2) Privacy Protection in Network Layer: Due to the network instability and the randomness of traffic loads in 6G systems, deep RL including A3C, DQN and DDPG has been applied in the network layer privacy protection to optimize the user location and the data caching policy and thus avoid information leakage and guarantee communication security. For example, the privacy protection scheme is designed in [204] to resist the malicious nodes at the network layer, which enables the MEC server to apply A3C with a global network and 20 worker agents in the MEC system to choose the privacy protection algorithm, the block interval of the blockchain, and the data flow selection based on the state, including the transmission rate of all the users, the user-server channel gain, the available computing resources of the MEC servers, the stake distribution, the user privacy level and the blockchain transaction size. The reward function is formulated with the transactional throughput of the blockchain system and the total energy consumption, but ignores the computational latency. To evaluate the privacy performance of the proposed scheme, a MEC system is simulated with Python 3.6 consisting of 20 users and 5 servers each having up to 8 GHz bandwidth and transmit power ranging from 0.1 to 2 W. The results show that the scheme improves the average throughput by 25.3%, and saves the average energy consumption by 16.1% compared with the privacy protection scheme without user data sharing.

In addition, the multi-access edge computing systems are vulnerable to malicious MEC servers in the network layer, in which the attackers can easily obtain the location privacy and usage pattern privacy of users during the computation tasks offloading process. Therefore, a deep RL based privacy preservation MEC system proposed in [205] replaces the two CNNs in the DQN architecture with two lightweight DNNs to protect user privacy and guarantee the QoS requirement for time-sensitive applications. More specifically, the state is formulated with the user location, the usage pattern privacy protection level and the channel gain. The system optimizes the protection level of the usage pattern privacy and the user location based on the state instead of the attack patterns and the user offloading strategies to maximize the reward including the user charging time and offloading time, without considering the privacy protection level. This scheme increases the reward by 30.0% compared with the local computing scheme in the MEC system equipped with a server and 10 user devices each of which has 915 MHz frequency and 5 W transmit power following the Rayleigh fading channel model.

The RL based privacy-preserving edge caching scheme in [206] combines the distributed DDPG algorithm with the federated learning technique to address the privacy leakage issue in MEC systems, assuming that each user has the same caching capacity.Different from the traditional DDPG, this algorithm optimizes the policy in parallel and updates the weights of the actor networks with a distributional critic network [210]. More specifically, the actor networks choose the caching policy of the current file, based on the state including the user request information and current cache to increase the global real-time cache hit rate. The distributional critic network of the MEC server updates the weights of the actor networks with N-step temporal difference error and the prioritized experience replay techniques to evaluate the chosen caching policy. Federated learning is used to predict the time-varying content popularity of the underlying file with the data privacy guarantee. On the other hand, each user also involves an actor network that optimizes their local caching decision based on the caching contents set of the user and the historical request information. In the simulated MEC system with a server and 10 users having 24 files to be cached, the proposed distributed DDPG scheme has a 64.8% higher cache hit rate than the randomized policy scheme.

In summary, most of the existing RL based cross-layer security and privacy protection schemes directly apply typical RL algorithms in the optimization of secure communication performance and privacy loss. Particularly, safe Q-learning has been applied in a cross-layer secure sensor-coordinator communication system against active eavesdropping. On the other hand, deep RL (such as PPO and DDPG) can be combined with safe exploration to further improve the cross-layer security performance with restricted QoS requirements.

XI. Future Research Directions

Interesting directions for the RL based 6G PHY-layer and cross-layer security include the security with partial observation, the safe exploration based security, privacy protection with federated RL, and multi-agent RL for 6G security.

A. 6G Security with Partial Observation

Existing RL based wireless security schemes have performance degradation under the partial and inaccurate observation of the state, which results from the inaccurate channel estimation and the delayed feedback over limited bandwidth in complicated and dynamic networks.

- **Partial observation:** The state estimation error such as the inaccurate channel estimation sometimes fails the learning of the security policy and yields security problems such as transmission failure or privacy leakage at both the PHY-layer and higher layers. This issue can be relieved by combining the model planning technique such as prioritized sweeping [211] with meta-learning algorithms such as the model-agnostic meta-learning (MAML) in [212]. More specifically, the model planning technique uses a virtual model to generate several simulated experiences in every time slot, and the meta-learning algorithms learn a large number of prior knowledge from similar security scenarios, in which both the simulated experiences and the prior knowledge are used to formulate an experience pool. Each wireless device exploits the previous security experiences (such as the authentication accuracy in [44]) from the experience pool. For example, the known information such as the location and transmit power of the other wireless devices can be exploited via the sampling training in MAML to address the state estimation error and select the power allocation strategy in 6G systems.
- **Observation delay:** The observation delay of the state and reward signals degrades the learning performance of the RL based security scheme. To this end, transfer learning such as fine-tuning based transfer learning in [213] can exploit the defense experiences in similar or simulated communication systems to initialize the learning parameters such as the CNN weights to help save the initial random exploration for more efficient security policy exploration.

B. Safe RL Based 6G Security

Wireless security applications are required to avoid the dangerous exploration of risky policies that cause serious security problems or privacy leakage. Safe RL algorithms, such as safe DQN, have been applied in anti-jamming communications and trust edge computing [55], [95], [98], [214]. In the future work, security metrics, such as the privacy level, the BER, and the authentication accuracy, should be formulated as the risk level in the exploration process, and the reward function has to incorporate the worse-case or constrained criterion, in order to further improve the PHY cross-layer security performance.

- **Criterion selection:** How to formulate an efficient security criterion in the reward function is critical for avoiding choosing risky policies, especially in the design of the PHY cross-layer security mechanisms. By incorporating the known network defense experiences and the expert guidance obtained from the teachers' advice or other similar scenarios, the mobile devices and BSs decide how to choose the criterion selection in 6G security applications such as anti-jamming communications [215].
- **Learning efficiency:** Existing safe RL based security schemes have to improve the optimization efficiency in the learning process to satisfy the latency requirements, especially the latency-sensitive applications such as real-time video games. Users can use inter-agent transfer learning to initialize the network parameters in the safe exploration and design DNNs to avoid the action or state quantization errors for higher learning efficiency.

C. Privacy Protection with Federated RL

Privacy issues in 6G systems will continue to exist, due to the data sharing between users and the third party, especially the selfish or malicious nodes, which can be addressed by federated learning [216]–[218]. As a promising decentralized machine learning technique, federated learning helps 6G systems avoid privacy leakage, improve the usage of the network bandwidth resources, and reduce the transmission latency for MEC, massive MIMO systems, IoT, and so on [219]. By applying federated learning, users or APs upload the learning or neural network parameters used in the computing model or policy selection rather than the user data to the central server for privacy protection [220].

However, the federated RL based privacy protection needs to address the following issues in practical 6G systems.

- **Inaccurate parameter updates and privacy leakage:** The information exploiting, data poisoning, model poisoning, and free-riding attacks launched by the malicious users can lead to inaccurate parameter updates of the central server and privacy leakage of other users in 6G systems. A potential solution is the blockchain-aided federated learning architecture, which records the parameter sharing among users without a third-party intermediary to suppress the malicious behaviors of users. In addition, DP can help protect user privacy by adding some "noise" such as Gaussian noise to the learning parameters before uploading them to the central server.
- **High communication overhead:** Due to the limited bandwidth, computational and energy resources of users, federated RL based privacy protection schemes require a large number of learning samples to converge. Thus, these schemes suffer from high communication and computational overhead per time slot in a large-scale decentralized and heterogenous 6G system. This issue can be well addressed by compression schemes such as sparsification and quantization that reduce the data exchange among users and the central server for less communication overhead. Besides, the users can apply imitation learning algorithms such as inverse RL that learn the reward function to provide additional learning samples and thus reduce the computational complexity.

D. Multi-agent RL for 6G Security

Wireless devices such as mobile devices and BSs that execute various tasks have to improve their learning efficiency in the selection of the PHY-layer and cross-layer security policy in large-scale 6G systems, which can be addressed by multi-agent RL. However, it is difficult for wireless devices to design a global learning optimization objective due to the independent goal of each device. The following two issues should be addressed before implementing multi-agent RL in 6G systems.

- **Exploration disaster:** The mobile device chooses its policy based on the observed network states as well as the policies of other agents and thus may result in a high-dimensional state space, yielding the exploration disaster. In addition, the devices with insufficient computational resources suffer from the high-dimensional discrete action set. A potential solution is exploiting the multi-agent communication mechanism such as DIAL that enables each device to explore the policies and states of the given neighboring devices instead of all the devices.
- **Privacy leakage:** In the multi-agent systems, mobile devices share their observations, policies and learning parameters such as the neural network weights with the other devices to improve the policy optimization efficiency but result in data or location privacy leakage. To address this issue, mobile devices can combine DP with multi-agent RL to determine the shared information and thus balance the privacy protection level and the learning efficiency.

XII. Summary

In this article, we have investigated the 6G PHY-layer attacks and shown that the NOMA, MEC, massive MIMO, mmWave, VLC, THz and IRS will be vulnerable to jamming, eavesdropping, DDoS, Sybil attacks, man-in-the-middle, selfish attacks and inference attacks. We reviewed the RL based PHY-security techniques for 6G systems and illustrated how to apply new RL algorithms to enhance the performance of the anti-jamming communications, secure communications, PHY-layer authentication, privacy-aware communications and location privacy protection with typical wireless techniques for 6G systems. Afterward, we summarized the RL based UAV security solutions at PHY-layer and discussed how the wireless devices apply RL to enhance the cross-layer security and privacy protection performance of 6G.

However, existing RL based security schemes will have severe performance degradation in 6G systems with large-scale heterogeneous networks to support real-time computation-intensive applications. Four major challenges for the RL based 6G security techniques include the inaccurate state and reward signals with long estimation latency, the risky state exploration, the security overhead, and the slow learning speed in large-scale networks. Promising solutions for robust and lightweight RL based 6G security are model planning, meta-learning, transfer learning, federated learning, safe RL and multi-agent RL.

References

[1] V.-L. Nguyen, P.-C. Lin, B.-C. Cheng, R.-H. Hwang, and Y.-D. Lin, “Security and privacy for 6G: A survey on prospective technologies and challenges,” *IEEE Commun. Surveys Tuts.*, vol. 23, no. 4, pp. 2384–2428, 4th Quart., 2021.

[2] R. Khan, P. Kumar, D. N. K. Jayakody, and M. Liyanage, “A survey on security and privacy of 5G technologies: Potential solutions, recent advancements, and future directions,” *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 196–248, 1st Quart., 2019.

[3] I. Ahmad, S. Shahabuddin, T. Kumar, J. Okwuibe, A. Gurtov, and M. Ylianttila, “Security for 5G and beyond,” *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3682–3722, 4th Quart., 2019.

[4] P. Yang, Y. Xiao, M. Xiao, and S. Li, “6G wireless communications: Vision and potential techniques,” *IEEE Netw.*, vol. 33, no. 4, pp. 70–75, Jul. 2019.

[5] F. Tang, Y. Kawamoto, N. Kato, and J. Liu, “Future intelligent and secure vehicular network toward 6G: Machine-learning approaches,” *Proc. IEEE*, vol. 108, no. 2, pp. 292–307, Feb. 2020.

[6] Y. Sun, J. Liu, J. Wang, Y. Cao, and N. Kato, “When machine learning meets privacy in 6G: A survey,” *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2694–2724, 4th Quart., 2020.

[7] S. Han *et al.*, “Artificial-intelligence-enabled air interface for 6G: Solutions, challenges, and standardization impacts,” *IEEE Commun. Mag.*, vol. 58, no. 10, pp. 73–79, Oct. 2020.

[8] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. sup Kwak, “Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges,” *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.

[9] T. X. Tran, A. Hajisami, P. Pandey, and D. Pompili, “Collaborative mobile edge computing in 5G networks: New paradigms, scenarios, and challenges,” *IEEE Commun. Mag.*, vol. 55, no. 4, pp. 54–61, Apr. 2017.

[10] H. Yang, W.-D. Zhong, C. Chen, and A. Alphones, "Integration of visible light communication and positioning within 5G networks for Internet of Things," *IEEE Netw.*, vol. 34, no. 5, pp. 134–140, Sept. 2020.

[11] K. M. S. Huq, S. A. Busari, J. Rodriguez, V. Frascolla, W. Bazzi, and D. C. Sicker, "Terahertz-enabled wireless system for beyond-5G ultra-fast networks: A brief survey," *IEEE Netw.*, vol. 33, no. 4, pp. 89–95, Jul. 2019.

[12] S. Gong *et al.*, "Toward smart wireless communications via intelligent reflecting surfaces: A contemporary survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2283–2314, 4th Quart., 2020.

[13] M. Giordani, M. Mezzavilla, and M. Zorzi, "Initial access in 5G mmWave cellular networks," *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 40–47, Nov. 2016.

[14] D. Kapetanovic, G. Zheng, and F. Rusek, "Physical layer security for massive MIMO: An overview on passive eavesdropping and active attacks," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 21–27, Jun. 2015.

[15] N. Wang, P. Wang, A. Alipour-Fanid, L. Jiao, and K. Zeng, "Physical-layer security of 5G wireless networks for IoT: Challenges and opportunities," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8169–8181, Oct. 2019.

[16] L. Xiao, X. Wan, C. Dai, X. Du, X. Chen, and M. Guizani, "Security in mobile edge caching with reinforcement learning," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 116–122, Jul. 2018.

[17] T. T. Do, E. Bjornson, E. G. Larsson, and S. M. Razavizadeh, "Jamming-resistant receivers for the massive MIMO uplink," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 1, pp. 210–223, Jan. 2018.

[18] Y. Cao *et al.*, "Secure transmission via beamforming optimization for NOMA networks," *IEEE Wireless Commun.*, vol. 27, no. 1, pp. 193–199, Feb. 2020.

[19] D. Darsena, G. Gelli, I. Iudice, and F. Verde, "Design and performance analysis of channel estimators under pilot spoofing attacks in multiple-antenna systems," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3255–3269, Apr. 2020.

[20] C. Ottaviani *et al.*, "Terahertz quantum cryptography," *IEEE JSAC*, vol. 38, no. 3, pp. 483–495, Mar. 2020.

[21] L. Sun and X. Tian, "Physical layer security in multi-antenna cellular systems: Joint optimization of feedback rate and power allocation," *IEEE Trans. Wireless Commun.*, vol. 1, no. 1, pp. 1–16, Mar. 2022.

[22] L. Xu, H. Xing, A. Nallanathan, Y. Yang, and T. Chai, "Security-aware cross-layer resource allocation for heterogeneous wireless networks," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1388–1399, Feb. 2019.

[23] M. Du, K. Wang, Y. Chen, X. Wang, and Y. Sun, "Big data privacy preserving in multi-access edge computing for heterogeneous Internet of Things," *IEEE Commun. Mag.*, vol. 56, no. 8, pp. 62–67, Aug. 2018.

[24] H. Zhao, M. Xu, Z. Zhong, and D. Wang, "A fast physical layer security-based location privacy parameter recommendation algorithm in 5G IoT," *China Commun.*, vol. 18, no. 8, pp. 75–84, Aug. 2021.

[25] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[26] U. Challita, H. Ryden, and H. Tullberg, "When machine learning meets wireless cellular networks: Deployment, challenges, and applications," *IEEE Commun. Mag.*, vol. 58, no. 6, pp. 12–18, Jun. 2020.

[27] M. Min *et al.*, "Learning-based privacy-aware offloading for healthcare IoT with energy harvesting," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4307–4316, Jun. 2019.

[28] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[29] V. Mnih *et al.*, "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 1928–1937, New York, NY, Jun. 2016.

[30] Lillicrap *et al.*, "Continuous control with deep reinforcement learning," in *Proc. Int. Conf. Learn. Representations (ICLR)*, pp. 1–14, San Juan, Puerto Rico, May 2016.

[31] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, Aug. 2017.

[32] L. Xiao, G. Sheng, S. Liu, H. Dai, M. Peng, and J. Song, "Deep reinforcement learning-enabled secure visible light communication against eavesdropping," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 6994–7005, Oct. 2019.

[33] T. T. Nguyen and V. J. Reddi, "Deep reinforcement learning for cyber security," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 1, no. 1, pp. 1–17, Nov. 2021.

[34] Z. Xiong, Y. Zhang, D. Niyato, R. Deng, P. Wang, and L.-C. Wang, "Deep reinforcement learning for mobile 5G and beyond: Fundamentals, applications, and challenges," *IEEE Veh. Technol. Mag.*, vol. 14, no. 2, pp. 44–52, Jun. 2019.

[35] Y. Dai, D. Xu, S. Maharjan, Z. Chen, Q. He, and Y. Zhang, "Blockchain and deep reinforcement learning empowered intelligent 5G beyond," *IEEE Netw.*, vol. 33, no. 3, pp. 10–17, May 2019.

[36] N. C. Luong *et al.*, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 4th Quart., 2019.

[37] L. U. Khan, W. Saad, D. Niyato, Z. Han, and C. S. Hong, "Digital-twin-enabled 6G: Vision, architectural trends, and future directions," *IEEE Commun. Mag.*, vol. 60, no. 1, pp. 74–80, Jan. 2022.

[38] P. Porambage, G. Gür, D. P. M. Osorio, M. Liyanage, A. Gurtov, and M. Ylianttila, "The roadmap to 6G security and privacy," *IEEE Open J. Commun. Society*, vol. 2, pp. 1094–1122, May 2021.

[39] K. Pelechrinis, M. Iliofotou, and S. V. Krishnamurthy, "Denial of service attacks in wireless networks: The case of jammers," *IEEE Commun. Surveys Tuts.*, vol. 13, no. 2, pp. 245–257, 2nd Quart., 2011.

[40] Q. Yan, H. Zeng, T. Jiang, M. Li, W. Lou, and Y. T. Hou, "Jamming resilient communication using MIMO interference cancellation," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 7, pp. 1486–1499, Jul. 2016.

[41] A. Mpitziopoulos, D. Gavalas, C. Konstantopoulos, and G. Pantziou, "A survey on jamming attacks and countermeasures in WSNs," *IEEE Commun. Surveys Tuts.*, vol. 11, no. 4, pp. 42–56, 4th Quart., 2009.

[42] L. Xiao, Y. Li, C. Dai, H. Dai, and H. V. Poor, "Reinforcement learning-based NOMA power allocation in the presence of smart jamming," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3377–3389, Apr. 2018.

[43] T. Erpek, Y. E. Sagduyu, and Y. Shi, "Deep learning for launching and mitigating wireless jamming attacks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 1, pp. 2–14, Mar. 2019.

[44] X. Lu, L. Xiao, T. Xu, Y. Zhao, Y. Tang, and W. Zhuang, "Reinforcement learning based PHY authentication for VANETs," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 3068–3079, Mar. 2020.

[45] A. K. Mishra, A. K. Tripathy, D. Puthal, and L. T. Yang, "Analytical model for sybil attack phases in Internet of Things," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 379–387, Feb. 2019.

[46] K. Zhang, X. Liang, R. Lu, and X. Shen, "Sybil attacks and their defenses in the Internet of Things," *IEEE Internet Things J.*, vol. 1, no. 5, pp. 372–383, Oct. 2014.

[47] A. Proano, L. Lazos, and M. Krunz, "Traffic decorrelation techniques for countering a global eavesdropper in WSNs," *IEEE Trans. Mobile Comput.*, vol. 16, no. 3, pp. 857–871, Mar. 2017.

[48] X. Sun, D. W. K. Ng, Z. Ding, Y. Xu, and Z. Zhong, "Physical layer security in UAV systems: Challenges and opportunities," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 40–47, Oct. 2019.

[49] A. Lu and G. Yang, "Stability analysis for cyber-physical systems under denial-of-service attacks," *IEEE Trans. Cybern.*, vol. 51, no. 11, pp. 5304–5313, Nov. 2021.

[50] V. L. Nguyen, P. Lin, and R. Hwang, "MECPASS: Distributed denial of service defense architecture for mobile networks," *IEEE Netw.*, vol. 32, no. 1, pp. 118–124, Jan. 2018.

[51] M. Conti, N. Dragoni, and V. Lesyk, "A survey of man in the middle attacks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 2027–2051, 3rd Quart., 2016.

[52] G. Oliva, S. Cioaba, and C. N. Hadjicostis, "Distributed calculation of edge-disjoint spanning trees for robustifying distributed algorithms against man-in-the-middle attacks," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 4, pp. 1646–1656, Dec. 2018.

[53] M. Jo, L. Han, D. Kim, and H. P. In, "Selfish attacks and detection in cognitive radio ad-hoc networks," *IEEE Netw.*, vol. 27, no. 3, pp. 46–50, May 2013.

[54] B. Jedari, F. Xia, and Z. Ning, "A survey on human-centric communications in non-cooperative wireless relay networks," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 914–944, 2nd Quart., 2018.

[55] L. Xiao *et al.*, "A reinforcement learning and blockchain-based trust mechanism for edge networks," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5460–5470, Sept. 2020.

[56] T. Wang, Z. Zheng, M. H. Rehmani, S. Yao, and Z. Huo, "Privacy preservation in big data from the communication perspective–A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 753–778, 1st Quart., 2019.

[57] A. Zhang and X. Lin, "Security-aware and privacy-preserving D2D communications in 5G," *IEEE Netw.*, vol. 31, no. 4, pp. 70–77, Jul. 2017.

[58] R. Lu, L. Zhang, J. Ni, and Y. Fang, "5G vehicle-to-everything services: Gearing up for security and privacy," *Proc. IEEE*, vol. 108, no. 2, pp. 373–389, Feb. 2020.

[59] M. Grissa, B. Hamdaoui, and A. A. Yavuz, "Location privacy in cognitive radio networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1726–1760, 3rd Quart., 2017.

[60] L. Xiao, X. Wan, X. Lu, Y. Zhang, and D. Wu, "IoT security techniques based on machine learning: How do IoT devices use AI to enhance security?," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 41–49, Sept. 2018.

[61] K. M. Thilina, K. W. Choi, N. Saquib, and E. Hossain, "Machine learning techniques for cooperative spectrum sensing in cognitive radio networks," *IEEE JSAC*, vol. 31, no. 11, pp. 2209–2221, Nov. 2013.

[62] G. Han, Y. He, J. Jiang, N. Wang, M. Guizani, and J. A. Ansere, "A synergetic trust model based on SVM in underwater acoustic sensor networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 11, pp. 11239–11247, Nov. 2019.

[63] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. Conf. Advances Neural Inf. Process. Syst. (NIPS)*, pp. 1057–1063, Denver, CO, Nov. 1999.

[64] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, May 1992.

[65] M. Bowling and M. Veloso, "Rational and convergent learning in stochastic games," in *Proc. Int. Joint Conf. Artificial Intell. (IJCAI)*, vol. 17, pp. 1021–1026, Seattle, WA, Aug. 2001.

[66] L. Xiao, X. Lu, D. Xu, Y. Tang, L. Wang, and W. Zhuang, "UAV relay in VANETs against smart jamming with reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4087–4097, May 2018.

[67] L. Xiao, Y. Li, G. Han, G. Liu, and W. Zhuang, "PHY-layer spoofing detection with reinforcement learning in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10037–10047, Dec. 2016.

[68] H. Li, S. Yu, X. Lu, L. Xiao, and L.-C. Wang, "Drone-aided network coding for secure wireless communications: A reinforcement learning approach," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, pp. 1–6, Madrid, Spain, Dec. 2021.

[69] B. Van Roy, D. P. Bertsekas, Y. Lee, and J. N. Tsitsiklis, "A neuro-dynamic programming approach to retailer inventory management," in *Proc. IEEE Conf. Decision and Control*, vol. 4, pp. 4052–4057, San Diego, CA, Dec. 1997.

[70] L. Xiao, T. Chen, C. Xie, H. Dai, and H. V. Poor, "Mobile crowdsensing games in vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1535–1545, Feb. 2018.

[71] G. A. Rummery and M. Niranjan, *On-line Q-learning using connectionist systems*. Cambridge, UK: University of Cambridge, Department of Engineering, 1994.

[72] H. Hasselt, "Double Q-learning," in *Proc. Conf. Advances Neural Inf. Process. Syst. (NIPS)*, vol. 23, pp. 2613–2621, Vancouver, Canada, Dec. 2010.

[73] A. Pritzel *et al.*, "Neural episodic control," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 2827–2836, Sydney, Australia, Aug. 2017.

[74] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. AAAI Conf. Artificial Intell.*, vol. 30, pp. 2094–2100, Feb. 2016.

[75] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 387–395, Beijing, China, Jun. 2014.

[76] Y. He, N. Zhao, and H. Yin, "Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 44–55, Jan. 2018.

[77] N. Wang, W. Li, A. Alipour-Fanid, M. Dabaghchian, and K. Zeng, "Compressed sensing-based pilot contamination attack detection for NOMA-IoT communications," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7764–7772, Aug. 2020.

[78] X. He, R. Jin, and H. Dai, "Physical-layer assisted secure offloading in mobile-edge computing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 4054–4066, Jun. 2020.

[79] L. Xiao, S. Hong, S. Xu, H. Yang, and X. Ji, "IRS-aided energy-efficient secure WBAN transmission based on deep reinforcement learning," *IEEE Trans. Commun.*, vol. 1, no. 1, pp. 1–13, Apr. 2022.

[80] Y. Yu, J. Tang, J. Huang, X. Zhang, D. K. C. So, and K.-K. Wong, "Multi-objective optimization for UAV-assisted wireless powered IoT networks based on extended DDPG algorithm," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 6361–6374, Sept. 2021.

[81] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE JSAC*, vol. 37, no. 10, pp. 2239–2250, Oct. 2019.

[82] C. He, Y. Hu, Y. Chen, and B. Zeng, "Joint power allocation and channel assignment for NOMA with deep reinforcement learning," *IEEE JSAC*, vol. 37, no. 10, pp. 2200–2210, Oct. 2019.

[83] H. Zhang, N. Yang, W. Huangfu, K. Long, and V. C. Leung, "Power control based on deep reinforcement learning for spectrum sharing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 4209–4219, Jun. 2020.

[84] L. Xiao, G. Sheng, X. Wan, W. Su, and P. Cheng, "Learning-based PHY-layer authentication for underwater sensor networks," *IEEE Commun. Lett.*, vol. 23, no. 1, pp. 60–63, Jan. 2019.

[85] A. Ferdowsi and W. Saad, "Deep learning for signal authentication and security in massive Internet-of-Things systems," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1371–1387, Feb. 2019.

[86] W. Liang, W. Huang, J. Long, K. Zhang, K.-C. Li, and D. Zhang, "Deep reinforcement learning for resource protection and real-time detection in IoT environment," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6392–6401, Jul. 2020.

[87] L. Buşoniu, R. Babuška, and B. De Schutter, "Multi-agent reinforcement learning: An overview," *Innovations Multi-agent Syst. Appl.-1*, pp. 183–221, 2010.

[88] L. Xiao, C. Xie, M. Min, and W. Zhuang, "User-centric view of unmanned aerial vehicle transmission against smart attacks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3420–3430, Apr. 2017.

[89] J. Foerster, I. A. Assael, N. De Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Proc. Conf. Advances Neural Inf. Process. Syst. (NIPS)*, pp. 2137–2145, Barcelona, Spain, Dec. 2016.

[90] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Conf. Advances Neural Inf. Process. Syst. (NIPS)*, vol. 30, Long Beach, CA, Dec. 2017.

[91] Z. Li and C. Guo, "Multi-agent deep reinforcement learning based spectrum allocation for D2D underlay communications," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1828–1840, Feb. 2020.

[92] H. Peng and X. Shen, "Multi-agent reinforcement learning based resource management in MEC-and UAV-assisted vehicular networks," *IEEE JSAC*, vol. 39, no. 1, pp. 131–141, Jan. 2021.

[93] L. Wang, K. Wang, C. Pan, W. Xu, N. Aslam, and L. Hanzo, "Multi-agent deep reinforcement learning based trajectory planning for multi-UAV assisted mobile edge computing," *IEEE Trans. Cogn.Commun. Netw.*, vol. 7, no. 1, Mar. 2021.

[94] N. Naderializadeh, J. J. Sydir, M. Simsek, and H. Nikopour, "Resource management in wireless networks via multi-agent deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 6, pp. 3507–3523, Jun. 2021.

[95] C. Dai, L. Xiao, X. Wan, and Y. Chen, "Reinforcement learning with safe exploration for network security," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP)*, pp. 3057–3061, Brighton, UK, May 2019.

[96] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward constrained policy optimization," in *Proc. Int. Conf. Learn. Representations (ICLR)*, New Orleans, LA, May 2019.

[97] S. Miryoosefi, K. Brantley, H. Daume III, M. Dudik, and R. E. Schapire, "Reinforcement learning with convex constraints," in *Proc. Conf. Advances Neural Inf. Process. Syst. (NIPS)*, pp. 14093–14102, Vancouver, Canada, Dec. 2019.

[98] L. Xiao, X. Lu, T. Xu, X. Wan, W. Ji, and Y. Zhang, "Reinforcement learning based mobile offloading for edge computing against jamming and interference," *IEEE Trans. Commun.*, vol. 68, no. 10, pp. 6114–6126, Oct. 2020.

[99] F. L. Da Silva and A. H. R. Costa, "A survey on transfer learning for multiagent reinforcement learning systems," *J. Artificial Intell. Research*, vol. 64, pp. 645–703, Mar. 2019.

[100] M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey," *J. Mach. Learn. Research*, vol. 10, no. 7, Sept. 2009.

[101] J. Wu, G. Li, S. Liu, and L. Lin, "Tree-structured policy based progressive reinforcement learning for temporally language grounding in video," in *Proc. AAAI Conf. Artificial Intel.*, vol. 34, pp. 12386–12393, New York, NY, Feb. 2020.

[102] S. Kelly and M. I. Heywood, "Knowledge transfer from keepaway soccer to half-field offense through program symbiosis: Building simple programs for a complex task," in *Proc. Annual Conf. Genetic Evol. Comput.*, pp. 1143–1150, Madrid, Spain, Jul. 2015.

[103] F. L. da Silva and A. H. R. Costa, "Accelerating multiagent reinforcement learning through transfer learning," in *Proc. AAAI Conf. Artificial Intell.*, pp. 5034–5035, San Francisc, CA, Feb. 2017.

[104] A. G. Barto and S. Mahadevan, "Recent advances in hierarchical reinforcement learning," *Discrete Event Dynamic Syst.*, vol. 13, no. 1, pp. 41–77, 2003.

[105] T. Kulkarni, K. Narasimhan, A. Saeedi, and J. Tenenbaum, "Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation," in *Proc. Conf. Advances Neural Inf. Process. Syst. (NIPS)*, pp. 3675–3683, Barcelona, Spain, Dec. 2016.

[106] L. Xiao, X. Lu, T. Xu, W. Zhuang, and H. Dai, "Reinforcement learning-based physical-layer authentication for controller area networks," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2535–2547, Feb. 2021.

[107] Vezhnevets *et al.*, "Feudal networks for hierarchical reinforcement learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 3540–3549, Sydney, Australia, Aug. 2017.

[108] J. Farah, E. P. Simon, P. Laly, and G. Delbarre, "Efficient combinations of NOMA with distributed antenna systems based on channel measurements for mitigating jamming attacks," *IEEE Syst. J.*, vol. 15, no. 2, pp. 2212–2221, Jun. 2020.

[109] H. Yang *et al.*, "Intelligent reflecting surface assisted anti-jamming communications: A fast reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1963–1974, Mar. 2021.

[110] X. Tang, D. Wang, R. Zhang, Z. Chu, and Z. Han, "Jamming mitigation via aerial reconfigurable intelligent surface: Passive beamforming and deployment optimization," *IEEE Trans. Veh. Technol.*, vol. 70, no. 6, pp. 6232–6237, Jun. 2021.

[111] J. Zhu, Z. Wang, Q. Li, H. Chen, and N. Ansari, "Mitigating intended jamming in mmWave MIMO by hybrid beamforming," *IEEE Wireless Commun. Lett.*, vol. 8, no. 6, pp. 1617–1620, Dec. 2019.

[112] Z. Xiao, B. Gao, S. Liu, and L. Xiao, "Learning based power control for mmWave massive MIMO against jamming," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, pp. 1–6, Abu Dhabi, UAE, Dec. 2018.

[113] H. Akhlaghpasand, E. Bjornson, and S. M. Razavizadeh, "Jamming-robust uplink transmission for spatially correlated massive MIMO systems," *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3495–3504, Jun. 2020.

[114] W. Zhang, Y. Hu, Y. Zhang, and D. Raychaudhuri, "SEGUE: Quality of service aware edge cloud service migration," in *Proc. IEEE Int. Conf. Cloud Comput. Technol. Sci. (CloudCom)*, pp. 344–351, Luxembourg City, Luxembourg, Dec. 2016.

[115] B. Akgun, M. Krunz, and O. Ozan Koyluoglu, "Vulnerabilities of massive MIMO systems to pilot contamination attacks," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 5, pp. 1251–1263, May 2019.

[116] Z. Shen, K. Xu, X. Xia, W. Xie, and D. Zhang, "Spatial sparsity based secure transmission strategy for massive MIMO systems against simultaneous jamming and eavesdropping," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3760–3774, Jun. 2020.

[117] D. Nguyen, P. Pathirana, M. Ding, and A. Seneviratne, "Secure computation offloading in blockchain based IoT networks with deep reinforcement learning," *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 4, pp. 3192–3208, Oct. 2021.

[118] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning based intelligent reflecting surface for secure wireless communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 375–388, Jan. 2021.

[119] C. Huang, G. Chen, and K.-K. Wong, "Multi-agent reinforcement learning-based buffer-aided relay selection in IRS-assisted secure cooperative networks," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4101–4112, Aug. 2021.

[120] T. K. Vu, M. Bennis, M. Debbah, M. Latva-aho, and C. S. Hong, "Ultra-reliable communication in 5G mmWave networks: A risk-sensitive approach," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 708–711, Apr. 2018.

[121] D. Zhao, H. Qin, B. Song, Y. Zhang, X. Du, and M. Guizani, "A reinforcement learning method for joint mode selection and power adaptation in the V2V communication network in 5G," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 2, pp. 452–463, Jun. 2020.

[122] X. Zhang and S. Sun, "Dynamic optimization for secure MIMO beamforming using large-scale reinforcement learning," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, pp. 1–6, Marrakesh, Morocco, Apr. 2019.

[123] L. Xu, A. Nallanathan, X. Pan, J. Yang, and W. Liao, "Security-aware resource allocation with delay constraint for NOMA-based cognitive radio network," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 2, pp. 366–376, Feb. 2018.

[124] Y. Zhou *et al.*, "Secure communications for UAV-enabled mobile edge computing systems," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 376–388, Jan. 2020.

[125] S. Han *et al.*, "Energy efficient secure computation offloading in NOMA-based mMTC networks for IoT," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5674–5690, Jun. 2019.

[126] Y. Chen, Y. Zhang, S. Maharjan, M. Alam, and T. Wu, "Deep learning for secure mobile edge computing in cyber-physical transportation systems," *IEEE Netw.*, vol. 33, no. 4, pp. 36–41, Jul. 2019.

[127] X. Wang, C. Chen, J. He, S. Zhu, and X. Guan, "Learning-based online transmission path selection for secure estimation in edge computing systems," *IEEE Trans. Ind. Informat.*, vol. 17, no. 5, pp. 3577–3587, May 2020.

[128] T. Bai, J. Wang, Y. Ren, and L. Hanzo, "Energy-efficient computation offloading for secure UAV-edge-computing systems," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 6074–6087, Jun. 2019.

[129] W. Sun, J. Liu, Y. Yue, and P. Wang, "Joint resource allocation and incentive design for blockchain-based mobile edge computing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 9, pp. 6050–6064, Sept. 2020.

[130] F. Wang *et al.*, "Optical jamming enhances the secrecy performance of the generalized space-shift-keying-aided visible-light downlink," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 4087–4102, Sept. 2018.

[131] S. Cho, G. Chen, and J. P. Coon, "Enhancement of physical layer security with simultaneous beamforming and jamming for visible light communication systems," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 10, pp. 2633–2648, Oct. 2019.

[132] V. Petrov, D. Moltchanov, J. M. Jornet, and Y. Koucheryavy, "Exploiting multipath terahertz communications for physical layer security in beyond 5G networks," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, pp. 865–872, Paris, France, Apr. 2019.

[133] W. Gao, Y. Chen, C. Han, and Z. Chen, "Distance-adaptive absorption peak modulation (DA-APM) for terahertz covert communications," *IEEE Trans. Wireless Commun.*, pp. 1–14, Nov. 2020.

[134] J. Ma *et al.*, "Security and eavesdropping in terahertz wireless links," *Nature*, vol. 563, no. 7729, pp. 89–93, Oct. 2018.

[135] W. Aman *et al.*, "Securing the insecure: A first-line-of-defense for nanoscale communication systems operating in THz band," *arXiv preprint arXiv:2007.06818*, Jul. 2020.

[136] H. Long *et al.*, "Reflections in the sky: Joint trajectory and passive beamforming design for secure UAV networks with reconfigurable intelligent surface," *arXiv preprint arXiv:2005.10559*, Jun. 2020.

[137] M. Cui, G. Zhang, and R. Zhang, "Secure wireless communication via intelligent reflecting surface," *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1410–1414, Oct. 2019.

[138] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 106–112, Jan. 2020.

[139] L. Dong and H. Wang, "Enhancing secure MIMO transmission via intelligent reflecting surface," *IEEE Tran. Wireless Commun.*, vol. 19, no. 11, pp. 7543–7556, Nov. 2020.

[140] X. Yu, D. Xu, Y. Sun, D. W. K. Ng, and R. Schober, "Robust and secure wireless communications via intelligent reflecting surfaces," *IEEE JSAC*, vol. 38, no. 11, pp. 2637–2652, Nov. 2020.

[141] S. Hong, C. Pan, H. Ren, K. Wang, and A. Nallanathan, "Artificial-noise-aided secure MIMO wireless communications via intelligent reflecting surface," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7851–7866, Dec. 2020.

[142] Y. Huang, J. Zhang, and M. Xiao, "Constant envelope hybrid precoding for directional millimeter-wave communications," *IEEE JSAC*, vol. 36, no. 4, pp. 845–859, Apr. 2018.

[143] Y. R. Ramadan, H. Minn, and A. S. Ibrahim, "Hybrid analog-digital precoding design for secrecy mmWave MISO-OFDM systems," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 5009–5026, Nov. 2017.

[144] Y. Hong, X. Jing, H. Gao, and Y. He, "Fixed region beamforming using frequency diverse subarray for secure mmWave wireless communications," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2706–2721, Jan. 2020.

[145] W. Wang and Z. Zheng, "Hybrid MIMO and phased-array directional modulation for physical layer security in mmWave wireless communications," *IEEE JSAC*, vol. 36, no. 7, pp. 1383–1396, Jul. 2018.

[146] S. Vuppala, Y. J. Tolossa, G. Kaddoum, and G. Abreu, "On the physical layer security analysis of hybrid millimeter wave networks," *IEEE Trans. Commun.*, vol. 66, no. 3, pp. 1139–1152, Mar. 2018.

[147] Y. Ju, H. Wang, Q. Pei, and H. Wang, "Physical layer security in millimeter wave DF relay systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 12, pp. 5719–5733, Dec. 2019.

[148] X. Sun, W. Yang, Y. Cai, L. Tao, Y. Liu, and Y. Huang, "Secure transmissions in wireless information and power transfer millimeter-wave

ultra-dense networks," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 7, pp. 1817–1829, Jul. 2019.

[149] M. Alageli, A. Ikhlef, F. Alsifiany, M. A. M. Abdullah, G. Chen, and J. Chambers, "Optimal downlink transmission for cell-free SWIPT massive MIMO systems with active eavesdropping," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1983–1998, Jan. 2020.

[150] J. Chen, X. Chen, W. H. Gerstacker, and D. W. K. Ng, "Resource allocation for a massive MIMO relay aided secure communication," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 8, pp. 1700–1711, Aug. 2016.

[151] D. Kudathanthirige, S. Timilsina, and G. A. Aruma Baduge, "Secure communication in relay-assisted massive MIMO downlink with active pilot attacks," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 11, pp. 2819–2833, Nov. 2019.

[152] W. Wang, N. Cheng, K. C. Teh, X. Lin, W. Zhuang, and X. Shen, "On countermeasures of pilot spoofing attack in massive MIMO systems: A double channel training based approach," *IEEE Trans. Veh. Technol.*, vol. 68, no. 7, pp. 6697–6708, Jul. 2019.

[153] N. Wang, L. Jiao, A. Alipour-Fanid, M. Dabaghchian, and K. Zeng, "Pilot contamination attack detection for NOMA in 5G mm-Wave massive MIMO networks," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1363–1378, Sept. 2019.

[154] M. Xiong, Y. Li, L. Gu, S. Pan, D. Zeng, and P. Li, "Reinforcement learning empowered IDPS for vehicular networks in edge computing," *IEEE Netw.*, vol. 34, no. 3, pp. 57–63, May 2020.

[155] S. Balakrishnan, S. Gupta, A. Bhuyan, P. Wang, D. Koutsonikolas, and Z. Sun, "Physical layer identification based on spatial–temporal beam features for millimeter-wave wireless networks," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1831–1845, Oct. 2020.

[156] P. Zhang, T. Taleb, X. Jiang, and B. Wu, "Physical layer authentication for massive MIMO systems with hardware impairments," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1563–1576, Mar. 2020.

[157] M. Du, K. Wang, Z. Xia, and Y. Zhang, "Differential privacy preserving of training model in wireless big data with edge computing," *IEEE Trans. Big Data*, vol. 6, no. 2, pp. 283–295, Jun. 2020.

[158] Y. Xiao, L. Xiao, X. Lu, H. Zhang, S. Yu, and H. V. Poor, "Deep reinforcement learning based user profile perturbation for privacy aware recommendation," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4560–4568, Mar. 2021.

[159] C. H. Liu, Q. Lin, and S. Wen, "Blockchain-enabled data collection and sharing for industrial IoT with deep reinforcement learning," *IEEE Trans. Ind. Informat.*, vol. 15, no. 6, pp. 3516–3526, Jun. 2019.

[160] Y. Liu, H. Wang, M. Peng, J. Guan, J. Xu, and Y. Wang, "DeePGA: A privacy-preserving data aggregation game in crowdsensing via deep reinforcement learning," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 4113–4127, May 2020.

[161] J. Feng, F. Richard Yu, Q. Pei, X. Chu, J. Du, and L. Zhu, "Cooperative computation offloading and resource allocation for blockchain-enabled mobile-edge computing: A deep reinforcement learning approach," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6214–6228, Jul. 2020.

[162] T. Zhao, F. Li, and L. He, "DRL-based joint resource allocation and device orchestration for hierarchical federated learning in NOMA-enabled industrial IoT," *IEEE Trans. Ind. Informat.*, vol. 1, no. 1, pp. 1–11, Apr. 2022.

[163] N. Xie, Q. Zhang, J. Chen, and H. Tan, "Privacy-preserving physical-layer authentication for non-orthogonal multiple access systems," *IEEE JSAC*, vol. 40, no. 4, pp. 1371–1385, Apr. 2022.

[164] J. Liu, C. Xiao, K. Cui, J. Han, X. Xu, and K. Ren, "Behavior privacy preserving in RF sensing," *IEEE Trans. Dependable Secure Comput.*, vol. 1, no. 1, pp. 1–12, Jan. 2022.

[165] Y. Wang, T. Gu, T. H. Luan, M. Lyu, and Y. Li, "HeartPrint: Exploring a heartbeat-based multi-user authentication with single mmWave radar," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 1–13, Aug. 2022.

[166] A. A. Nasir, H. D. Tuan, and T. Q. Duong, "Fractional time exploitation for serving IoT users with guaranteed QoS by 5G spectrum," *IEEE Commun. Mag.*, vol. 56, no. 10, pp. 128–133, Oct. 2018.

[167] Y.-S. Jeon, M. M. Amiri, J. Li, and H. V. Poor, "A compressive sensing approach for federated learning over massive MIMO communication systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1990–2004, Mar. 2021.

[168] Y. Sun, Q. Liu, X. Chen, and X. Du, "An adaptive authenticated data structure with privacy-preserving for big data stream in cloud," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3295–3310, Apr. 2020.

[169] Y. Yu *et al.*, "Identity-based remote data integrity checking with perfect data privacy preserving for cloud storage," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 4, pp. 767–778, Apr. 2017.

[170] Z. Xiao, X. Fu, and R. S. M. Goh, "Data privacy-preserving automation architecture for industrial data exchange in smart cities," *IEEE Trans. Ind. Informat.*, vol. 14, no. 6, pp. 2780–2791, Jun. 2018.

[171] Z. Huang, R. Hu, Y. Guo, E. Chan-Tin, and Y. Gong, "DP-ADMM: ADMM-based distributed learning with differential privacy," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1002–1012, Jul. 2019.

[172] J. Xiong *et al.*, "A personalized privacy protection framework for mobile crowdsensing in IIoT," *IEEE Trans. Ind. Informat.*, vol. 16, no. 6, pp. 4231–4241, Jun. 2020.

[173] R. Guerraoui, A.-M. Kermarrec, R. Patra, and M. Taziki, "D2P: Distance-based differential privacy in recommenders," in *Proc. Int. Conf. Very Large Data Bases (VLDB) Endowment*, vol. 8, pp. 862–873, Kohala Coast, HI, Sept. 2015.

[174] D. C. Nguyen, P. N. Pathirana, M. Ding, and A. Seneviratne, "Privacy-preserved task offloading in mobile blockchain with deep reinforcement learning," *IEEE Trans. Netw. Service Manag.*, vol. 17, no. 4, pp. 2536–2549, Dec. 2020.

[175] X. He, R. Jin, and H. Dai, "Deep PDS-learning for privacy-aware offloading in MEC-enabled IoT," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4547–4555, Jun. 2019.

[176] Y. Dai, D. Xu, K. Zhang, S. Maharjan, and Y. Zhang, "Deep reinforcement learning and permissioned blockchain for content caching in vehicular edge computing and networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 4312–4324, Apr. 2020.

[177] M. Min, W. Wang, L. Xiao, Y. Xiao, and Z. Han, "Reinforcement learning-based sensitive semantic location privacy protection for VANETs," *China Commun.*, vol. 18, no. 6, pp. 244–260, Jun. 2021.

[178] Q. Pan, J. Wu, X. Zheng, W. Yang, and J. Li, "Differential privacy and IRS empowered intelligent energy harvesting for 6G Internet of Things," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 1–14, Aug. 2021.

[179] M. A. Ul Alam, M. M. Rahman, and J. Q. Widberg, "PALMAR: Towards adaptive multi-inhabitant activity recognition in point-cloud technology," in *Proc. IEEE Conf. Computer Commun. (INFOCOM)*, pp. 1–10, Vancouver, Canada, May 2021.

[180] Z. Zhang, R. Yang, X. Zhang, C. Li, Y. Huang, and L. Yang, "Backdoor federated learning-based mmWave beam selection," *IEEE Trans. Commun.*, vol. 1, no. 1, pp. 1–16, Aug. 2022.

[181] J. Xu, X. Wang, P. Zhu, and X. You, "Privacy-preserving channel estimation in cell-free hybrid massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 6, pp. 3815–3830, Jun. 2021.

[182] Z. Montazeri, A. Houmansadr, and H. Pishro-Nik, "Achieving perfect location privacy in wireless devices using anonymization," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 11, pp. 2683–2698, Nov. 2017.

[183] S. Zou, J. Xi, H. Wang, and G. Xu, "CrowdBLPS: A blockchain-based location-privacy-preserving mobile crowdsensing system," *IEEE Trans. Ind. Informat.*, vol. 16, no. 6, pp. 4206–4218, Jun. 2020.

[184] X. He, R. Jin, and H. Dai, "Leveraging spatial diversity for privacy-aware location-based services in mobile networks," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 6, pp. 1524–1534, Jun. 2018.

[185] W. Zhang, M. Li, R. Tandon, and H. Li, "Online location trace privacy: An information theoretic approach," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 1, pp. 235–250, Jun. 2019.

[186] C. Yin, J. Xi, R. Sun, and J. Wang, "Location privacy protection based on differential privacy strategy for big data in industrial Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 14, no. 8, pp. 3628–3636, Aug. 2018.

[187] Z. Li, Y. Lu, X. Li, Z. Wang, W. Qiao, and Y. Liu, "UAV networks against multiple maneuvering smart jamming with knowledge-based reinforcement learning," *IEEE Internet Things J.*, vol. 8, no. 15, pp. 12289–12310, Aug. 2021.

[188] L. Xiao, Y. Ding, J. Huang, S. Liu, Y. Tang, and H. Dai, "UAV anti-jamming video transmissions with QoE guarantee: A reinforcement learning-based approach," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 5933–5947, Sept. 2021.

[189] X. Lu, J. Jie, Z. Lin, L. Xiao, J. Li, and Y. Zhang, "Reinforcement learning based energy efficient robot relay for unmanned aerial vehicles against smart jamming," *Sci. China Inf. Sci.*, vol. 65, no. 1, pp. 1–13, Jan. 2022.

[190] Y. Zhang, Z. Mou, F. Gao, J. Jiang, R. Ding, and Z. Han, "UAV-enabled secure communications by multi-agent deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11599–11611, Oct. 2020.

[191] Z. Zhang, Q. Zhang, J. Miao, F. R. Yu, F. Fu, J. Du, and T. Wu, "Energy-efficient secure video streaming in UAV-enabled wireless networks: A safe-DQN approach," *IEEE Trans. Green Commun. Netw.*, vol. 5, no. 4, pp. 1892–1905, Dec. 2021.

[192] Y. Wang, Z. Su, N. Zhang, and A. Benslimane, "Learning in the air: Secure federated learning for UAV-assisted crowdsensing," *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 2, pp. 1055–1069, Apr. 2021.

[193] J. Tao, T. Han, and R. Li, "Deep-reinforcement-learning-based intrusion detection in aerial computing networks," *IEEE Netw.*, vol. 35, no. 4, pp. 66–72, Jul. 2021.

[194] Y. Nie, J. Zhao, F. Gao, and F. R. Yu, "Semi-distributed resource management in UAV-aided MEC systems: A multi-agent federated reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 13162–13173, Dec. 2021.

[195] H. Yang, J. Zhao, Z. Xiong, K.-Y. Lam, S. Sun, and L. Xiao, "Privacy-preserving federated learning for UAV-enabled networks: Learning-based joint scheduling and resource management," *IEEE JSAC*, vol. 39, no. 10, pp. 3144–3159, Oct. 2021.

[196] A. Fotouhi *et al.*, "Survey on UAV cellular communications: Practical aspects, standardization advancements, regulation, and security challenges," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3417–3442, 4th Quart., 2019.

[197] X. Lu, L. Xiao, C. Dai, and H. Dai, "UAV-aided cellular communications with deep reinforcement learning against jamming," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 48–53, Aug. 2020.

[198] L. Xiao, H. Li, S. Yu, Y. Zhang, L.-C. Wang, and S. Ma, "Reinforcement learning based network coding for drone-aided secure wireless communications," *IEEE Trans. Commun.*, vol. 70, no. 9, pp. 5975–5988, Sept. 2022.

[199] D. Zhang, F. R. Yu, R. Yang, and L. Zhu, "Software-defined vehicular networks with trust management: A deep reinforcement learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 2, pp. 1400–1414, Feb. 2022.

[200] Y. Li, L. Zhu, H. Wang, F. R. Yu, and S. Liu, "A cross-layer defense scheme for edge intelligence-enabled CBTC systems against MitM attacks," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 4, pp. 2286–2298, Apr. 2020.

[201] B. Gu, L. Gao, X. Wang, Y. Qu, J. Jin, and S. Yu, "Privacy on the edge: Customizable privacy-preserving context sharing in hierarchical edge computing," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 4, pp. 2298–2309, Aug. 2020.

[202] Z. Xue, P. Zhou, Z. Xu, X. Wang, Y. Xie, X. Ding, and S. Wen, "A resource-constrained and privacy-preserving edge-computing-enabled clinical decision system: A federated reinforcement learning approach," *IEEE Internet Things J.*, vol. 8, no. 11, pp. 9122–9138, Feb. 2021.

[203] K. Tang, N. Kan, J. Zou, C. Li, X. Fu, M. Hong, and H. Xiong, "Multi-user adaptive video delivery over wireless networks: A physical layer resource-aware deep reinforcement learning approach," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 798–815, Mar. 2021.

[204] L. Liu *et al.*, "Blockchain-enabled secure data sharing scheme in mobile-edge computing: An asynchronous advantage actor–critic learning approach," *IEEE Internet Things J.*, vol. 8, no. 4, pp. 2342–2353, Dec. 2020.

[205] P. Zhao, J. Tao, L. Kangjie, G. Zhang, and F. Gao, "Deep reinforcement learning-based joint optimization of delay and privacy in multiple-user MEC systems," *IEEE Trans. Cloud Comput.*, vol. 1, no. 1, pp. 1–13, Jan. 2022.

[206] S. Liu, C. Zheng, Y. Huang, and T. Q. Quek, "Distributed reinforcement learning for privacy-preserving dynamic edge caching," *IEEE JSAC*, vol. 40, no. 3, pp. 749–760, Mar. 2022.

[207] L. Zhu, Y. Li, F. R. Yu, B. Ning, T. Tang, and X. Wang, "Cross-layer defense methods for jamming-resistant CBTC systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 11, pp. 7266–7278, Nov. 2021.

[208] T. Zhao, L. He, X. Huang, and F. Li, "QoE-driven secure video transmission in cloud-edge collaborative networks," *IEEE Trans. Veh. Technol.*, vol. 71, no. 1, pp. 681–696, Jan. 2022.

[209] D. Xu, K. Yu, and J. A. Ritcey, "Cross-layer device authentication with quantum encryption for 5G enabled IIoT in industry 4.0," *IEEE Trans. Ind. Informat.*, vol. 1, no. 1, pp. 1–11, Nov. 2021.

[210] G. Barth-Maron *et al.*, "Distributed distributional deterministic policy gradients," in *Proc. Int. Conf. Learn. Representations (ICLR)*, pp. 1–16, Vancouver, Canada, Apr. 2018.

[211] A. W. Moore and C. G. Atkeson, "Prioritized sweeping: Reinforcement learning with less data and less time," *Mach. learn.*, vol. 13, no. 1, pp. 103–130, 1993.

[212] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 1126–1135, Sydney, Australia, Aug. 2017.

[213] S. Gamrian and Y. Goldberg, "Transfer learning for related reinforcement learning tasks via image-to-image translation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 2063–2072, Long Beach, CA, Jun. 2019.

[214] X. Lu, L. Xiao, G. Niu, X. Ji, and Q. Wang, "Safe exploration in wireless security: A safe reinforcement learning algorithm with hierarchical structure," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 732–743, Feb. 2022.

[215] J. Huang, F. Wu, D. Precup, and Y. Cai, "Learning safe policies with expert guidance," in *Proc. Conf. Advances Neural Inf. Process. Syst. (NIPS)*, pp. 9105–9114, Montreal, Canada, Dec. 2018.

[216] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.

[217] J. Zhang, B. Chen, X. Cheng, H. T. T. Binh, and S. Yu, "PoisonGAN: Generative poisoning attacks against federated learning in edge computing systems," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3310–3322, Mar. 2021.

[218] H. Yang *et al.*, "Lead federated neuromorphic learning for wireless edge artificial intelligence," *Nature Commun.*, vol. 13, no. 1, pp. 1–12, Jul. 2022.

[219] W. Y. B. Lim *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.

[220] T. T. Vu, D. T. Ngo, N. H. Tran, H. Q. Ngo, M. N. Dao, and R. H. Middleton, "Cell-free massive MIMO for wireless federated learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6377–6392, Oct. 2020.

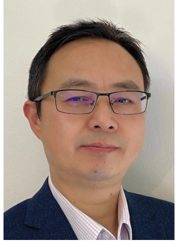
Xiaozhen Lu (Member, IEEE) received the B.S. degree in communication engineering from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2017, and the Ph.D. degree in communication and information systems from Xiamen University, Xiamen, China, in 2021. She is currently an Associate Professor with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. She was a recipient of the Best Student Paper Award for ML4CS 2019, and the Excellent Paper Award for CWSN 2020. Her research interests include reinforcement learning, network security, and wireless communications.

Liang Xiao (Senior Member, IEEE) received the B.S. degree in communication engineering from the Nanjing University of Posts and Telecommunications, China, in 2000, the M.S. degree in electrical engineering from Tsinghua University, China, in 2003, and the Ph.D. degree in electrical engineering from Rutgers University, NJ, USA, in 2009. She was a Visiting Professor with Princeton University, Virginia Tech, and the University of Maryland, College Park. She is currently a Professor with the Department of Information and Communication Engineering, Xiamen University, Xiamen, China. She was a recipient of the Best Paper Award for 2016 INFOCOM Big Security WS and 2017 ICC. She has served as an Associate Editor for IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and a Guest Editor for IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING.

Pengmin Li received the B.S. degree in communication engineering from Xiamen University, Xiamen, China, in 2020, where she is currently pursuing the M.S. degree with the Department of Information and Communication Engineering. Her research interests include network security and wireless communications.

Xiangyang Ji (Member, IEEE) received the B.S. and the M.S. degrees from the Harbin Institute of Technology, Harbin, China, in 1999 and 2001, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He joined Tsinghua University, Beijing, in 2008, where he is currently a Professor at the Department of Automation, School of Informtion Science and Technology. He has authored more than 100 refereed conference and journal papers. His current research interests include signal processing, computer vision and computational photography.

Chenren Xu (Senior Member, IEEE) received the PhD degree from WINLAB, Rutgers University, Brunswick, New Jersey. He was a postdoctoral fellow in Carnegie Mellon University, Pittsburgh, Pennsylvania and visiting scholar in AT&T Shannon Labs and Microsoft Research. He is an assistant professor with the Department of Computer Science and Technology as well as an affiliated member of CECA at Peking University, China, where he directs Software-hardware Orchestrated ARchitecture (SOAR) Lab since 2015. His research interests span wireless, networking and system. He is the recipient of Alibaba DAMO Academy Young Fellow and CCF-Intel Young Faculty Awards.



Shui Yu (Senior Member, IEEE) obtained his PhD from Deakin University, Australia, in 2004. He currently is a Professor of School of Computer Science, University of Technology Sydney, Australia. Dr Yu's research interest includes Big Data, Security and Privacy, Networking, and Mathematical Modelling. He has published four monographs and edited two books, more than 500 technical papers, including top journals and top conferences, such as IEEE TPDS, TC, TIFS, TMC, TKDE, TETC, ToN, and INFOCOM. His h-index is 66. Dr Yu initiated the research field of networking for big data in 2013, and his research outputs have been widely adopted by industrial systems, such as Amazon cloud security. He is currently serving a number of prestigious editorial boards, including IEEE Communications Surveys and Tutorials (Area Editor), IEEE Communications Magazine, IEEE Internet of Things Journal, and so on. He served as a Distinguished Lecturer of IEEE Communications Society (2018-2021). He is a Distinguished Visitor of IEEE Computer Society, a voting member of IEEE ComSoc Educational Services board, and an elected member of Board of Governor of IEEE Vehicular Technology Society.

Weihua Zhuang (Fellow, IEEE) is a University Professor and a Tier I Canada Research Chair in Wireless Communication Networks at University of Waterloo, Canada. Her research focuses on network architecture, algorithms and protocols, and service provisioning in future communication systems. She is the recipient of 2021 Women's Distinguished Career Award from IEEE Vehicular Technology Society, 2021 Technical Contribution Award in Cognitive Networks from IEEE Communications Society, 2021 R.A. Fessenden Award from IEEE Canada, and 2021 Award of Merit from the Federation of Chinese Canadian Professionals in Ontario. She was the Editor-in-Chief of the IEEE Transactions on Vehicular Technology from 2007 to 2013, General Co-Chair of 2021 IEEE/CIC International Conference on Communications in China (ICCC), Technical Program Chair/Co-Chair of 2017/2016 IEEE VTC Fall, Technical Program Symposia Chair of 2011 IEEE Globecom, and an IEEE Communications Society Distinguished Lecturer from 2008 to 2011. She is an elected member of the Board of Governors and the Executive Vice President of the IEEE Vehicular Technology Society. Dr. Zhuang is a Fellow of the IEEE, Royal Society of Canada, Canadian Academy of Engineering, and Engineering Institute of Canada.