# Applications of Reinforcement Learning in Wireless Communication Channel Estimation: Current Research, Achievements, and Innovative Frontiers

## 1. Introduction: The Rise of Reinforcement Learning in Channel Estimation

### 1.1 The Evolving Landscape of Channel Estimation in Wireless Systems

In wireless communication systems, channel estimation plays a crucial role, as its accuracy directly impacts key performance indicators such as data transmission rates, bit error rates, and signal coverage.1 Traditional channel estimation algorithms, like Least Squares (LS) and Linear Minimum Mean Square Error (LMMSE), often struggle to accurately capture channel characteristics in complex communication environments, such as those with significant multipath effects, rapidly changing channel states, or strong interference, leading to decreased estimation accuracy.1 With the evolution of next-generation wireless networks (NextG, 5G, 6G), application scenarios are becoming increasingly diverse, including Massive Multiple-Input Multiple-Output (Massive MIMO), millimeter Wave (mmWave) communications, Intelligent Reflecting Surfaces (IRS), and Vehicle-to-Everything (V2X) communications. These all impose stricter requirements on the accuracy and adaptability of channel estimation.3

This technological shift is not merely an incremental improvement of existing algorithms but a response to the fundamental limitations of traditional methods in increasingly complex and dynamic next-generation network environments. Traditional methods often rely on statistical channel models like Gaussian or Rayleigh, which may no longer be applicable in highly dynamic, non-stationary, or interference-laden real-world scenarios.1 While supervised deep learning can learn these complex mappings, it requires a vast amount of labeled data for every possible channel condition, which is practically unfeasible.

### 1.2 Reinforcement Learning: A Paradigm Shift in Adaptive Channel Estimation

Reinforcement Learning (RL) offers a new paradigm for adaptive channel estimation. It is a data-driven, model-free (or model-based) approach where an agent learns to make optimal decisions through interaction with the environment.7 This characteristic makes it particularly suitable for dynamically changing, unknown channel environments where traditional mathematical models may be inaccurate or difficult to obtain.7 RL can continuously optimize decisions in dynamic channel environments, thereby improving the robustness of channel estimation.1

The application of RL in channel estimation signifies a move towards a higher level of autonomy and intelligence in communication systems, where the system can self-optimize physical layer functions. An agent making channel estimation decisions via RL, such as selecting pilot symbols 12, configuring Intelligent Reflecting Surfaces (IRS) 1, or jointly optimizing transceivers 10, represents a greater degree of automation than pre-programmed algorithms or offline-trained supervised models. This aligns with the broader vision of AI-native 6G networks.4

### 1.3 Advantages of Reinforcement Learning Over Traditional and Conventional Machine Learning Methods

Compared to traditional channel estimation algorithms and conventional machine learning methods (like supervised learning), RL exhibits several advantages:

* **Adaptability:** The core strength of RL lies in its ability to adapt to dynamically changing environments through continuous learning and policy adjustment.1 Unlike supervised learning, which requires large labeled datasets (often difficult to obtain for all channel conditions), RL learns from experience and rewards.
* **Model-Free Capability:** Many RL algorithms do not require an exact mathematical model of the channel, which is particularly important in scenarios where the channel model is complex or unknown, especially in end-to-end (E2E) systems or when non-linear characteristics are present.7 While this "model-free" nature provides robustness to unmodeled dynamics, it can also lead to higher sample complexity and challenges in ensuring convergence to a global optimum compared to model-based methods or supervised learning with accurate labels. Reward function design becomes critical, and ensuring it guides the agent towards the desired channel estimation objective is non-trivial.7
* **Optimization of Long-Term Performance:** RL aims to maximize cumulative rewards, thereby formulating strategies that are optimal in the long run, rather than just making greedy choices for immediate gain.15
* **Handling Non-Linearities:** Deep Reinforcement Learning (DRL), which combines RL with deep neural networks, can effectively handle non-linearities and dynamic fluctuations in complex channel environments through automatic feature extraction.1

## 2. Core Reinforcement Learning Methodologies for Channel Estimation

Reinforcement learning offers a diverse algorithmic toolkit for channel estimation. These methods can be broadly categorized into value-based and policy-gradient methods, the latter often combined with Actor-Critic architectures. In recent years, Deep Reinforcement Learning (DRL), by introducing deep neural networks as function approximators, has significantly expanded the capability of these methods to handle high-dimensional and complex problems, which is crucial for the complex channel state representations and continuous control parameter spaces in modern communication systems.8

### 2.1 Value-Based Methods: Fundamentals and Applications

Value-based methods aim to learn a value function that estimates the "worth" of taking a certain action in a particular state or being in a certain state.

* **Q-Learning and Deep Q-Networks (DQN): Principles and Channel Estimation Strategies**
  + **Q-Learning:** A model-free, off-policy RL algorithm that learns a state-action value function (Q-value), representing the expected cumulative reward for taking a specific action in a given state.7 It typically uses a Q-table to store these values. In channel estimation, Q-learning has been applied in various scenarios. For instance, in Multi-Input Single-Output Non-Orthogonal Multiple Access (MISO-NOMA) systems, a Q-learning agent at the base station can predict downlink channel coefficients to maximize system sum-rate and reduce estimation loss.15 In such applications, the state can include current user channel conditions, actions are predictions or adjustments to channel coefficients, and rewards are based on system sum-rate.15 Another application is successive denoising in MIMO OFDM systems, where Q-learning identifies unreliable channel estimates by recognizing channel curvature and applies geometry-based denoising actions to update channel estimates. Its state is the quantized channel estimate, actions involve selecting subcarriers for denoising, and rewards reflect the reduction in Mean Squared Error (MSE).17
  + **Deep Q-Network (DQN):** An extension of Q-learning that uses a deep neural network to approximate the Q-value function, enabling it to handle large-scale continuous state spaces.7 DQN often employs techniques like experience replay and target networks to enhance learning stability. In channel coefficient estimation for NOMA systems, DQN is used to predict channel parameters with the goal of maximizing downlink sum-rate.18 DQN overcomes the limitations of traditional Q-learning in terms of Q-table storage.18 Furthermore, improved algorithms like Double DQN (DDQN) can mitigate overestimation bias in DQN.19
* Performance Insights and Inherent Limitations  
  Q-learning and its deep learning-based variants have achieved significant performance improvements in specific channel estimation tasks. For example, they have outperformed traditional MMSE methods in terms of Bit Error Rate (BER), outage probability, and system sum-rate.15 DQN can effectively handle high-dimensional state spaces.19  
  However, these methods also have inherent limitations. Q-learning can suffer from an impractically large Q-table when dealing with complex problems, leading to storage and computational difficulties.15 While DQN addresses the state space issue, it can suffer from overestimation bias 7 and is sensitive to hyperparameter selection.19 Additionally, both methods are typically suited for discrete action spaces, which is a limitation for optimizing many continuous parameters in channel estimation (e.g., precise phase shifts).7 Computational complexity and convergence time are also concerns.15

### 2.2 Policy Gradient and Actor-Critic Methods: Advanced Strategies

Policy gradient methods directly learn a policy function that maps states to actions. Actor-Critic methods are a popular implementation of policy gradient methods, comprising an "actor" network that learns the policy and a "critic" network that evaluates how good the policy is.

* **Deep Deterministic Policy Gradient (DDPG):** DDPG is an off-policy, model-free Actor-Critic algorithm designed for continuous action spaces. The actor network learns a deterministic policy (mapping states to specific actions), and the critic network learns a Q-value function (evaluating the actions chosen by the actor).1 Applications of DDPG in channel estimation include: using features extracted by a CNN-GRU model as state input to optimize IRS reflection parameters, where the DDPG agent continuously learns and adjusts IRS configurations to improve signal quality and coverage.1 In end-to-end communication systems, DDPG is used to jointly train the transmitter and receiver under unknown channel conditions without prior Channel State Information (CSI), using the receiver's loss as a reward signal.10
* **Proximal Policy Optimization (PPO):** PPO is an on-policy Actor-Critic algorithm known for its learning stability and sample efficiency. It uses a clipped surrogate objective function to limit the magnitude of policy updates, thereby avoiding excessively large policy changes.7 Application examples include: optimizing RIS phase shifts, UAV altitude, and communication scheduling by considering Signal-to-Noise Ratio (SNR) and previous RIS phase shifts 7; optimizing Restricted Access Window (RAW) parameters in IEEE 802.11ah IoT networks to improve throughput 10; and optimizing channel sampling patterns in Integrated Sensing and Communication (ISAC) systems.28
* **Twin Delayed DDPG (TD3):** TD3 is an extension of DDPG aimed at addressing DDPG's overestimation bias and improving its stability. It achieves these improvements through techniques such as clipped double Q-learning, delayed policy updates, and target policy smoothing.7 Applications of TD3 include: optimizing the RIS phase-shift matrix in RIS-assisted networks by considering the channel's response to local information to maximize the total achievable finite blocklength rate 7; and jointly optimizing satellite beamforming, UAV positioning, and power allocation in integrated satellite-UAV-RIS networks.30 An enhanced TD3 variant, EE-DDPG, is also proposed in.21
* Performance Insights and Inherent Limitations  
  Actor-Critic methods like DDPG and PPO are well-suited for continuous action spaces, which is crucial for many parameters in channel estimation (e.g., beamforming vectors, phase shifts).1 DDPG can achieve significant performance in end-to-end systems without channel knowledge.10 PPO offers a stable and efficient learning process.16 TD3 improves upon DDPG's stability and overestimation issues.7  
  However, these advanced strategies also face challenges. DDPG can suffer from overestimation bias and has high computational complexity due to multiple neural networks.7 PPO, being an on-policy algorithm, can be sample inefficient. TD3 adds further complexity on top of DDPG.7 Moreover, hyperparameter tuning remains a challenge for these algorithms.19

The choice of RL algorithm is closely tied to the action space characteristics (discrete vs. continuous) of the channel estimation problem. For instance, selecting detected data symbols in data-aided channel estimation 12 naturally fits the discrete action space of Q-learning/DQN. Conversely, optimizing continuous parameters like IRS phase shifts 1, beamforming vectors, or power levels 21 necessitates policy gradient or Actor-Critic methods like DDPG, PPO, or TD3, which are designed for continuous action spaces.7

The evolution of DRL algorithms, such as from DQN to DDQN and from DDPG to TD3, reflects the ongoing efforts by researchers to address fundamental RL challenges like overestimation bias, learning stability, and sample efficiency, which are critical for achieving reliable and practical channel estimation. Overestimation bias (addressed by DDQN 19 and TD3 7) can lead to suboptimal policies in channel estimation, degrading performance. Unstable learning processes make training outcomes unreliable. These improvements are not just academic; they are essential for deploying RL in real communication systems where consistent performance is paramount. The development of algorithms like PPO 7 also highlights the search for a balance between performance, stability, and ease of implementation.

Furthermore, a recurring, critical, and often subtly handled aspect is "reward function design," which significantly impacts the success of RL in channel estimation. Various rewards are mentioned in the literature, such as minimizing channel estimation error 12, maximizing sum-rate 15, improving signal quality 1, or using receiver-side loss.10 The effectiveness of an RL agent is directly tied to how well this reward signal aligns with the true objective of channel estimation. A poorly designed reward can lead the agent to learn unintended or suboptimal behaviors, even if the RL algorithm itself is powerful. This is a key "knob" that researchers must carefully tune.

Table 1 below summarizes the key RL algorithms used for channel estimation and their critical features.

**Table 1: Summary of Key Reinforcement Learning Algorithms for Channel Estimation**

| **Algorithm** | **Type** | **Key Features (e.g., Discrete/Continuous Action Space, On/Off-Policy, Model-Free)** | **Typical MDP Components for Channel Estimation (State; Action; Reward)** | **Primary Application Areas in Channel Estimation (e.g., Data-Aided CE, IRS Optimization, E2E Learning, Denoising)** | **Key References** |
| --- | --- | --- | --- | --- | --- |
| Q-Learning | Value-Based | Discrete action space, off-policy, model-free | **State:** Received signal, pilot data, historical channel estimates, quantized channel estimates; **Action:** Select pilot/data symbols, adjust channel coefficient values, select subcarriers for denoising; **Reward:** MSE/NMSE reduction, BER improvement, sum-rate maximization | Data-aided CE, channel prediction, MIMO OFDM denoising | 7 |
| DQN | Value-Based | Discrete action space (typically), off-policy, model-free, uses DNN to approximate Q-function | **State:** High-dimensional channel features, received signal; **Action:** Select pilot/data symbols, predict channel parameters; **Reward:** Sum-rate maximization, SINR enhancement | NOMA system channel estimation, large state-space problems | 7 |
| DDPG | Actor-Critic | Continuous action space, off-policy, model-free | **State:** CSI features (e.g., from CNN-GRU), receiver-side loss information; **Action:** Adjust IRS phases, set beamforming vectors, adjust transmitter parameters; **Reward:** Signal quality improvement, coverage enhancement, BER/BLER reduction | IRS optimization, End-to-End (E2E) learning, continuous parameter control | 1 |
| PPO | Actor-Critic | Discrete or continuous action space, on-policy, model-free, clipped surrogate objective function | **State:** SNR, previous RIS phase shifts, network state; **Action:** Adjust IRS phase shifts/UAV altitude/scheduling, set RAW parameters, select channel sampling patterns; **Reward:** Age of Information minimization, throughput enhancement, SINR enhancement | RIS optimization, IoT network parameter optimization, ISAC systems | 7 |
| TD3 | Actor-Critic | Continuous action space, off-policy, model-free, addresses DDPG overestimation | **State:** Local CSI, system state; **Action:** Optimize RIS phase-shift matrix, adjust beamforming/power/rate splitting; **Reward:** Finite blocklength rate maximization, energy efficiency maximization | RIS optimization, integrated network resource management | 7 |

## 3. Reinforcement Learning in Action: Channel Estimation Across Diverse Communication Scenarios

The adaptive and learning capabilities of RL enable it to address unique channel estimation challenges across various communication scenarios. From intelligent reflecting surfaces to highly dynamic mobile environments, and emerging mmWave and IoT applications, RL is being actively explored to enhance the performance and robustness of channel estimation. The characteristics of these application scenarios, such as channel dynamics, hardware limitations, and system objectives, directly dictate the formulation of the Markov Decision Process (MDP) – including the definition of states, actions, and rewards – and the choice of the corresponding RL algorithm.

### 3.1 Intelligent Reflecting Surfaces (IRS) / Reconfigurable Intelligent Surfaces (RIS): RL-Driven Optimization

* **Problem Context:** IRS/RIS technology intelligently manipulates the wireless propagation environment using a large number of low-cost passive reflecting units to enhance received signal quality or suppress interference. However, the introduction of IRS/RIS complicates channel estimation, often involving cascaded user-IRS-base station channels that are high-dimensional and difficult to estimate and track directly.1 Traditional methods struggle to adapt to the dynamic configuration needs of IRS.2
* **RL Application:** Deep Reinforcement Learning, particularly algorithms like DDPG and PPO, is widely used to learn the optimal phase shift configurations for IRS/RIS reflecting units.1 The RL agent receives current state information (e.g., extracted channel features, previous phase shift configurations, SNR of the received signal) and outputs actions (i.e., adjustments to phase shifts). The reward function is typically designed based on metrics like received signal quality, Signal-to-Interference-plus-Noise Ratio (SINR), or system sum-rate.
* **Typical Case:** Reference 1 proposes an intelligent channel estimation model that combines features extracted by a Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU) as state input and uses the DDPG algorithm to optimize IRS configuration. This model continuously learns and adjusts IRS configuration strategies in dynamic channel environments, significantly improving channel estimation accuracy and robustness over traditional LS and LMMSE methods. The survey in 7 also notes that DDPG can be used to optimize RIS phase shifts based on CSI, and it also investigates the application of DQN, TD3, and PPO in RIS.7
* **Key Achievements:** Optimizing IRS/RIS via RL can significantly enhance SINR, improve channel capacity, and achieve adaptive optimization in dynamic environments, thereby boosting overall communication system performance.1

### 3.2 Tackling Dynamics: RL in Time-Varying and Doubly-Dispersive Channels

* **Problem Context:** Wireless channels are inherently time-varying, especially in high-speed mobile scenarios where channel characteristics change rapidly. Doubly-dispersive channels, whose characteristics vary in both time and frequency domains, are even more complex. These rapid variations pose significant challenges to traditional channel estimation algorithms, which often assume the channel is quasi-static over a period.3
* **RL Application:** RL is used to select detected data symbols as additional pilots in time-varying channels to aid channel estimation, thereby better adapting to channel changes.12 By modeling the symbol selection problem as an MDP, RL algorithms can compute the optimal symbol selection policy. Mechanisms like state element refinement and sliding windows help track the dynamic changes of the channel.12
* **Typical Case:** Reference 12 proposes an RL-aided channel estimator for time-varying MIMO systems that outperforms traditional estimators by effectively capturing channel variations. This work models the channel using a first-order Gauss-Markov process and adjusts the weights of data symbols according to channel changes. Reference 14 also emphasizes similar methods for time-varying MIMO systems and notes that deriving the optimal solution is very difficult due to computational complexity and channel dynamics.
* **Key Achievements:** RL methods can more effectively track temporal channel variations, outperforming traditional estimation algorithms in dynamic scenarios.12

### 3.3 Enhancing Multi-Antenna Systems: RL in MIMO and Massive MIMO Channel Estimation

* **Problem Context:** MIMO systems use multiple antennas to improve spectral efficiency and link reliability, but this requires accurate acquisition of high-dimensional CSI between antennas. Massive MIMO systems scale the number of antennas to very large numbers, further exacerbating challenges related to pilot overhead and estimation complexity.5 Moreover, the Hybrid Analog-Digital (HAD) architecture often employed in Massive MIMO also increases the difficulty of channel estimation, as the receiver can only observe a few linear combinations of the original signal.32
* **RL Application:** RL is used for data-aided channel estimation by selectively utilizing detected symbols as additional pilots to reduce estimation error.13 To address computational demands, researchers have proposed low-complexity RL algorithms, such as reducing processing complexity by dividing data blocks into sub-blocks.13 Q-learning algorithms are also used for successive denoising of channel estimates in MIMO OFDM systems.17 In Massive MIMO, attention mechanisms can be combined with deep learning to enable the model to focus on relevant channel features, thereby improving estimation accuracy.32
* **Typical Case:** Reference 13 proposes a low-complexity RL channel estimator for MIMO systems that utilizes detected symbols and defines the problem as an MDP to minimize MSE, while also introducing a discount factor. Reference 17 uses Q-learning for successive denoising in MIMO OFDM systems based on channel curvature.
* **Key Achievements:** The application of RL significantly reduces MSE/NMSE and BER, improves performance with limited pilots, and exhibits good robustness in time-varying channels.13

### 3.4 Beyond Explicit Models: RL in End-to-End (E2E) Communication Systems

* **Problem Context:** Traditional end-to-end learning systems often assume known prior channel information or a differentiable channel model for backpropagation, which is impractical in many real-world unknown or non-differentiable channel scenarios.10 RL, particularly DDPG, offers a solution for such "black-box" channel estimation problems.
* **RL Application:** DDPG-based end-to-end systems allow joint training of transmitter and receiver neural networks without prior channel knowledge.10 In this architecture, the transmitter acts as the DDPG agent, and the receiver's loss information is used as the reward signal. Channel estimation is implicitly handled by the neural networks, and the system learns to communicate effectively over the unknown channel.
* **Typical Case:** References 10 and 10 describe a DDPG-E2E system that utilizes loss feedback from the receiver to optimize the transmitter, achieving superior Block Error Rate (BLER) performance over solutions requiring channel models in complex wireless channels (e.g., Rician, Rayleigh, and 3GPP channels), especially when handling large block length data. Reference 11 provides a technical explanation and critical analysis of this DDPG-E2E approach.
* **Key Achievements:** Achieved joint optimization of transmitter and receiver without explicit channel models and obtained better BLER performance over complex wireless channels.10

### 3.5 RL Exploration in Emerging Frontiers: mmWave, IoT, V2X, and Underwater Channel Estimation

* **Millimeter Wave (mmWave):** mmWave communications face challenges like high path loss and sensitivity to blockage. RL can be used to jointly optimize beamforming and ADC thresholds for low-resolution ADCs in mmWave MIMO systems.34 In RIS-assisted mmWave systems, statistical CSI (S-CSI) can be utilized for optimization.35
* **Internet of Things (IoT) / Backscatter Communication:** Low power consumption is a key requirement for IoT applications. RL algorithms (e.g., DQN, DDQN, DDPG, PPO) can adaptively adjust transmission parameters (e.g., modulation scheme, transmit power) and enhance channel estimation in backscatter systems by adapting to SNR and interference variations.19 For example, PPO has been used for channel hopping or channel blacklisting mechanisms to avoid interference and improve communication reliability.26
* **Vehicle-to-Everything (V2X):** V2X communication environments are highly dynamic with rapidly changing channels. Deep Reinforcement Learning (including Single-Agent RL (SARL) and Multi-Agent RL (MARL)) is used for spectrum sharing and resource allocation, decisions that implicitly rely on an understanding of channel states.9 DRL can also optimize beamforming based on sensing state information, thereby reducing reliance on pilots for channel estimation.39
* **Underwater Communication (UWC):** Underwater channel environments are complex, time-varying, and often suffer from severe multipath effects. RL (e.g., Q-learning, DQN) is explored for adaptive modulation based on (potentially outdated) CSI 41, and for optimizing transmission parameters.7 While research directly applying RL for underwater channel estimation is not prominent in the provided literature, deep learning methods (like neural networks) have been applied 43, and RL could complement these approaches.

In these emerging areas, while RL sometimes handles channel estimation implicitly (e.g., in E2E systems or resource allocation tasks), its potential is immense. A notable trend is that, compared to fields like MIMO or IRS, RL frameworks specifically for *direct* channel parameter estimation might be less mature in some emerging areas (like V2X or UWC). This suggests that developing more explicit RL methods tailored for the channel estimation task itself in these domains could be a fruitful avenue for innovation.

The successful application of RL across diverse communication scenarios heralds a future of highly adaptive communication systems capable of operating effectively even in poorly understood or rapidly changing environments, reducing the need for manual system tuning and extensive pre-deployment modeling. This is crucial for future networks (6G and beyond), which are expected to be far more complex and heterogeneous than current ones.4

Table 2 summarizes the application of RL for channel estimation problems in different communication scenarios.

**Table 2: Applications of Reinforcement Learning for Channel Estimation in Different Communication Scenarios**

| **Communication Scenario** | **Channel Estimation Problem Addressed (e.g., Cascaded channel estimation, Dynamic channel tracking, High-dimensional CSI acquisition, Implicit CE in E2E, Beamforming/ADC optimization, Interference avoidance, Robust CE in harsh environments)** | **RL Methods Employed (e.g., DDPG, PPO, Q-Learning, DQN, Actor-Critic)** | **Key Performance Results/Improvements (e.g., NMSE/BER reduction, Sum-rate/SINR increase, Faster adaptation, CSI-free operation)** | **Challenges in this Scenario (e.g., Dimensionality, Dynamics, Pilot overhead, Power constraints)** | **Key References** |
| --- | --- | --- | --- | --- | --- |
| IRS/RIS | Cascaded channel estimation, Dynamic phase shift optimization | DDPG, PPO, DQN, TD3 | NMSE reduction, SINR/Sum-rate increase, Dynamic adaptation | High dimensionality, Cascaded channel, Rapid changes | 1 |
| Time-Varying/Doubly-Dispersive Channels | Dynamic channel tracking, Data symbol-aided estimation | RL (general), Q-Learning | Effective capture of channel variations, Outperforms traditional estimators | Rapid time-variance, Computational complexity | 3 |
| MIMO/Massive MIMO | High-dimensional CSI acquisition, Pilot overhead reduction, Estimation in HAD architecture, Channel denoising | RL (low-complexity), Q-Learning, DL with attention mechanism | MSE/NMSE/BER reduction, Improved performance with limited pilots, Robustness in time-varying channels | Extremely high dimensionality, Pilot contamination, Computational complexity | 13 |
| End-to-End (E2E) Systems | Implicit channel estimation and joint transceiver optimization without prior CSI | DDPG | Joint optimization without explicit channel model, Improved BLER in complex channels | Black-box channel, Gradient propagation, Large block length processing | 10 |
| Millimeter Wave (mmWave) | Joint optimization of beamforming and low-resolution ADCs, Robust estimation under high path loss | RL (general), Policy Gradient | Performance close to exhaustive search, Robust to noisy CSI | High path loss, Blockage sensitivity, Hardware limitations | 34 |
| IoT/Backscatter | Adaptive transmission in low-power settings, Interference avoidance, Channel parameter adaptation | DQN, DDQN, DDPG, PPO | Adaptation to SNR/interference changes, Improved spectral efficiency | Extremely low power constraints, Simple device computational capability | 19 |
| Vehicle-to-Everything (V2X) | Channel state understanding and resource allocation in highly dynamic environments, Sensing-based beamforming | DRL (SARL, MARL), Actor-Critic | Improved spectrum sharing efficiency, Reduced pilot dependency | Extreme dynamics, Rapid channel changes, Low latency requirements | 9 |
| Underwater Communication (UWC) | Adaptive modulation and parameter optimization in harsh multipath and time-varying channels | Q-Learning, DQN | Improved adaptive modulation performance with outdated CSI | Extreme channel conditions, Severe multipath, Bandwidth limitations | 7 |

## 4. Significant Achievements and Performance Benchmarks of RL Channel Estimation

The application of RL to channel estimation has yielded a series of encouraging results, particularly in terms of performance improvement and environmental adaptability, demonstrating clear advantages over traditional methods and other machine learning techniques. These achievements are typically quantified through Key Performance Indicators (KPIs) and compared against existing benchmarks.

### 4.1 Quantified Improvements Over Traditional Estimators (LS, LMMSE) and Other ML Techniques

Studies have shown that RL-based frameworks, especially DRL, achieve significantly lower channel estimation errors (e.g., Normalized Mean Squared Error - NMSE, Mean Squared Error - MSE) than traditional LS and LMMSE methods in various communication scenarios.1 For example, an IRS-assisted channel estimation model using DDPG exhibited far lower channel estimation error than LS and LMMSE methods.1 In MISO-NOMA systems, Q-learning-based channel prediction not only improved BER but also achieved higher sum-rates than MMSE and even deep learning-based LSTM methods.15 Furthermore, an RL-aided denoising method for MIMO OFDM systems performed close to ideal LMMSE estimation (which requires perfect channel statistics) and significantly outperformed practical LS estimation.17

This performance superiority, especially in dynamic or complex scenarios (like IRS, time-varying channels), suggests that RL's ability to learn non-linear mappings and adapt to changing statistics is a key differentiator. LS/LMMSE typically rely on linear models and second-order statistics 1, struggling to effectively capture complex channel phenomena (e.g., multipath, non-linear distortions, fast fading). DRL, with neural networks as function approximators, can learn these complex relationships directly from data 1, leading to the superior performance demonstrated in studies like.1

### 4.2 Key Performance Indicators (KPIs) and Achieved Gains

* **Bit Error Rate (BER) / Block Error Rate (BLER):** RL methods have been shown to improve BER.13 For instance, DDPG-E2E systems achieved better BLER performance under various channel conditions.10
* **Mean Squared Error (MSE) / Normalized Mean Squared Error (NMSE):** Significant reductions in MSE/NMSE are commonly reported achievements in RL channel estimation.1
* **Signal-to-Interference-plus-Noise Ratio (SINR) / Sum-Rate:** In systems like RIS and NOMA, improvements in SINR and sum-rate are often optimization targets for RL agents, and practical performance gains have been achieved.7

### 4.3 Validated Adaptability and Robustness in Complex and Dynamic Environments

RL enables models to continuously optimize decisions in dynamic channel environments, thereby enhancing the robustness of channel estimation.1 RL-aided estimators can effectively capture channel variations in time-varying MIMO systems.12 DDPG-E2E systems have demonstrated good robustness under complex channel models like Rician, Rayleigh, and 3GPP.10 An important feature is that RL methods can operate without prior channel knowledge, adapting to unknown communication environments.7

The frequently reported achievements of "adaptability" and "robustness" in the literature are direct manifestations of RL's learning paradigm – interacting with the environment and optimizing based on feedback (rewards). Unlike fixed algorithms or supervised models trained on static datasets, RL agents continuously refine their policies based on the outcomes of their actions in the current (and potentially changing) environment.1 This inherent feedback loop allows them to adapt to unforeseen channel conditions or dynamics not present in the initial training set, leading to the observed robustness.1

These achievements of RL in channel estimation are paving the way for more autonomous and self-optimizing wireless networks, reducing the need for manual intervention and pre-defined models for every possible scenario. The ability of RL to achieve high performance in scenarios where accurate channel models are difficult to obtain (e.g., E2E systems 10, IRS 1) or where channels are highly dynamic (12) points towards systems that can self-configure and self-optimize. This aligns with the goals of future intelligent networks that can manage complex resources and adapt to unpredictable conditions with minimal human supervision.

However, it is important to note that while RL generally outperforms traditional methods, the reported gains are often specific to the problem formulation (MDP, reward function) and the DRL architecture used. Generalizing these gains to all possible channel estimation tasks or achieving "plug-and-play" RL solutions remains an open challenge. Successful instances (e.g., 1) involve carefully designed states, actions, and rewards tailored to the specific problem (e.g., IRS phase control, symbol selection, denoising). A DDPG agent excelling in IRS control 1 might not achieve optimal performance in, for example, V2X channel tracking without significant redesign of its learning framework. This highlights that while the RL *paradigm* itself is powerful, its *application* still requires domain-specific expertise.

## 5. Overcoming Hurdles: Challenges and Limitations of RL in Channel Estimation Applications

Despite the immense potential demonstrated by RL in channel estimation, several key challenges and inherent limitations need to be addressed before its widespread deployment in practical communication systems. These challenges span computational overhead, scalability, data efficiency, robustness, and security.

### 5.1 Computational Demands: Complexity and Latency Issues

RL algorithms, particularly Deep Reinforcement Learning (DRL) incorporating deep neural networks, can be computationally intensive during both training and inference phases.7 Computing the optimal policy can entail considerable complexity and latency.12 For instance, DDPG algorithms have significant computational overhead due to multiple neural networks 7, while traditional Q-learning's Q-table can become unmanageable with large state spaces.15 These computational demands pose a barrier to practical deployment, especially in real-time systems or on resource-constrained terminal devices.11

### 5.2 Scalability and Practicality in Large-Scale Deployments

Scaling RL solutions to systems with a large number of users, antennas (as in Massive MIMO), or IRS elements remains a formidable challenge.7 As system size increases, the state and action spaces can grow exponentially, making the learning process intractable for many RL algorithms.13 The challenges of computational complexity, scalability, and data efficiency are often intertwined. A highly complex DRL model might offer better performance but will also demand more data and computational power, hindering its scalability. Thus, researchers must trade off performance against practical feasibility, especially in real-world systems. Efforts to develop "lightweight" DRL 46 directly target this cluster of challenges.

### 5.3 The Critical Role of MDP Formulation and Reward Engineering

Defining appropriate states, actions, and especially the reward function is crucial for RL success but is non-trivial.7 The reward function must accurately reflect the objective of channel estimation. Inadequate MDP design can lead to suboptimal policies or slow convergence. The "black-box" nature of DRL models, while beneficial for unknown environments, also adds to the difficulty of reward engineering and ensuring generalization. The lack of a clear analytical model for the DRL agent's decision-making process makes it harder to design perfect reward functions.7 It also makes it more difficult to predict agent behavior in novel, out-of-distribution scenarios, impacting generalization.11 This contrasts with model-based optimization, where behavior is more predictable.

### 5.4 Data Efficiency, Sample Complexity, and Generalization Capability

RL algorithms can suffer from low sample efficiency, meaning they require a large number of interactions with the environment to learn an effective policy.11 Ensuring that models can generalize to channel conditions not encountered during training or to environments different from the training setup is critical.3 Furthermore, offline training of DRL models often requires large-scale datasets 46, which can be difficult to obtain in some scenarios.

### 5.5 Security Considerations: Vulnerability to Adversarial Manipulation

Deep learning-based channel estimation models, including those using RL, are vulnerable to adversarial machine learning attacks.4 Attackers could craft perturbations to deceive the receiver with incorrect channel estimation results, posing a serious security risk.5 As ML/RL models become increasingly integral to communication systems, security is an emerging and critical challenge. If channel estimation, a fundamental physical layer process, can be maliciously manipulated through attacks on the RL agent, the entire communication link can be compromised. This moves beyond performance optimization into the realm of system trustworthiness and resilience, key themes for next-generation networks.4

Overcoming these challenges is crucial for transitioning RL-based channel estimation from academic research to widespread practical deployment. The current limitations define the major research directions needed for the field to mature. Solutions in these areas (e.g., efficient algorithms, transfer/meta-learning for data efficiency, robust RL for security) will enable RL to become a standard tool in the communications toolkit.

## 6. The Path Forward: Innovative Frontiers and Research Directions for RL in Channel Estimation

Despite significant progress, there is ample room for innovation and numerous research directions to explore in RL for channel estimation. Future research trends primarily focus on enhancing the practicality, adaptability, efficiency, trustworthiness, and integration of RL with other communication modules to meet the growing complexity and diverse demands of next-generation wireless communication systems.

### 6.1 Advanced Learning Paradigms for Enhanced Adaptability and Efficiency

* **Meta-Reinforcement Learning (Meta-RL): Enabling Rapid Adaptation to New Channel Environments**
  + **Concept:** Meta-RL, or "learning to learn," aims to train agents that can quickly adapt to new tasks with minimal data, such as different channel environments, user mobility patterns, or new Quality of Service (QoS) requirements.48 This is crucial for handling non-stationary channels.
  + **Application:** Models can be trained to find optimal initial parameters that allow for rapid convergence to new channel conditions.48 Model-Agnostic Meta-Learning (MAML) is a prominent algorithm.48 For example, 48 combines MAML with DQN for optimizing flight trajectory and scheduling policies in UAV communications to minimize a combination of Age of Information (AoI) and transmission power. Reference 49 introduces GMA, a context-based meta-RL method with Mixture of Experts (MoE) to improve representation learning for rapid adaptation of MAC protocols.
  + **Potential:** Reduce training overhead in new environments, improve generalization capabilities.48 Meta-RL can also tune hyperparameters of RL solutions for fast adaptation.48
  + **Related Research:** 48 (MAML with DQN for UAV AoI/power optimization), 49 (GMA, context-based meta-RL with MoE for MAC protocol adaptation), 54 (Meta-learning for downlink channel prediction, adapting to new environments with few data), 52 (MAML for time-varying OFDM channel estimation, rapid adaptation with few samples), 50 (Offline meta-RL).
* **Transfer Learning (TL): Knowledge Transfer Across Communication Tasks and Scenarios**
  + **Concept:** Utilizing knowledge gained from source tasks/environments to improve learning in target tasks/environments, thereby reducing reliance on large amounts of data in the target domain.8
  + **Application:** Models can be pre-trained on data from various environments and then fine-tuned with limited data from a new environment.54 For example, 54 proposes a direct transfer algorithm in FDD massive MIMO systems, where a model is trained on data from all previous environments and then fine-tuned with a small amount of labeled data in a new environment to predict downlink CSI from uplink CSI. Transfer learning can also be used to initialize local model parameters in Heterogeneous Federated Learning (HFL).55
  + **Potential:** Accelerate the learning process, improve performance in data-scarce scenarios, and adapt more effectively to new system configurations or channel statistics.
  + **Related Research:** 54 (Direct transfer and meta-learning for FDD massive MIMO channel prediction), 55 (TL for enhancing HFL initialization for channel estimation in RIS-assisted cell-free MIMO systems).
* **Federated and Distributed Reinforcement Learning: Towards Privacy-Preserving and Scalable Channel Estimation**
  + **Concept:** Collaboratively training RL models across multiple devices (users, base stations) without sharing raw local data, thereby enhancing privacy.4 Distributed RL allows agents to learn policies based on local information while contributing to a global objective.
  + **Application:** Federated learning can be used for channel estimation in RIS-assisted cell-free MIMO systems, where users form coalitions and use distributed DRL (e.g., Qmix-based methods) for decision-making.55
  + **Potential:** Improve scalability, reduce communication overhead of centralized training, protect user data privacy, and support on-device learning.
  + **Related Research:** 55 (HFL with coalition formation and DRL for channel estimation), 7 (Mentions FL/SL as future research directions for RIS), 4 (Distributed learning-based channel estimation models for post-5G systems). The need for rapid adaptation to highly dynamic and diverse 6G scenarios (e.g., varying services, user densities, channel conditions) is a strong driver for meta-RL and transfer learning research. Traditional RL, and even DRL, trained for a specific scenario, can suffer significant performance degradation when conditions change markedly. Retraining from scratch is often too time-consuming or data-intensive.49 Meta-RL 48 and TL 54 provide mechanisms to "learn to adapt" or "transfer knowledge," which is crucial for the flexibility envisioned in future networks.

### 6.2 Designing Lightweight and Computationally Efficient DRL Algorithms

* **Problem:** The high computational cost of DRL hinders its deployment on resource-constrained devices.11
* **Methods:**
  + Developing simpler neural network architectures, e.g., lightweight CNNs combined with Transformers.46
  + Exploring methods like infinite-width neural networks/Neural Tangent Kernels (CNTK) that may not require large datasets or extensive training.46 These methods leverage the property that the training dynamics of infinite-width networks can be described by closed-form equations to estimate missing channel responses using only known values at pilot locations.
  + Performing algorithmic-level optimizations to reduce the complexity of Q-learning or Actor-Critic methods, such as the sub-block partitioning strategy proposed in.13
  + Applying model compression techniques like pruning, quantization, or knowledge distillation to DRL models.
* **Potential:** Enable on-device RL channel estimation, reduce latency, and make RL more practical in real-time applications.
* **Related Research:** 46 (CNTK for channel estimation, reducing data and computational needs), 13 (Low-complexity RL via sub-block partitioning), 16 (PPO for IoT RAW optimization, implying a focus on efficiency). Reference 20 mentions DQN can handle large state spaces, but computational efficiency remains a concern. The push for lightweight DRL 46 and federated/distributed RL 4 are complementary. Lightweight models are easier to deploy on edge devices, a prerequisite for many federated learning scenarios. Federated learning involves training models on local devices.55 If these devices are resource-constrained (common in IoT, V2X), the DRL models must be computationally inexpensive. Thus, advances in lightweight DRL directly facilitate more effective and widespread federated learning applications in channel estimation.

### 6.3 Building Robust RL Agents for Non-Stationary, High-Interference, and Adversarial Environments

* **Problem:** Real wireless channels are often non-stationary, subject to unpredictable interference, and potentially face adversarial attacks.1
* **Methods:**
  + Employing RL algorithms inherently designed for dynamic environments, such as DDPG used in 1 for IRS adjustment under complex conditions, and RL methods for time-varying channels in.12
  + Developing RL agents with explicit robustness objectives or constraints.
  + Combining RL with interference suppression techniques, e.g., DRL-optimized Rate Splitting Multiple Access (RSMA) to mitigate interference in.58
  + Investigating secure RL methods to defend against adversarial attacks on channel estimators.4 This may involve adversarial training or robust reward function design.
* **Potential:** Ensure reliable communication performance even in challenging and hostile wireless environments.
* **Related Research:** 1 (DDPG for dynamic IRS adjustment), 58 (DRL for RSMA precoding to mitigate interference), 4 (Security concerns and need for robust models).

### 6.4 Towards Explainable AI (XAI): Enhancing Interpretability of RL-Driven Channel Estimators

* **Problem:** DRL models, especially deep neural networks, are often treated as "black boxes," making their decision-making processes difficult to understand.11 This lack of interpretability can hinder trust and debugging.
* **Methods:**
  + Developing XAI techniques tailored for RL agents in communication systems.
  + Visualizing learned policies or value functions, e.g., attention maps for Massive MIMO in.32
  + Designing RL architectures that are inherently more interpretable.
  + Utilizing learned world models to understand agent predictions, as in 59 based on the agent's internal state or world model predictions of future actions.
* **Potential:** Enhance trust in RL-based systems, facilitate debugging and improvement, and provide insights into learned channel characteristics.
* **Related Research:** 32 (Visualizing attention maps for interpretability), 11 (Lack of interpretability cited as a limitation of DDPG-E2E), 59 (Predicting agent future behavior based on internal state or world model for understanding).

### 6.5 Holistic Optimization: RL for Joint Channel Estimation and Cross-Layer Design

* **Concept:** Beyond optimizing channel estimation in isolation, RL can be used to jointly optimize channel estimation with other communication tasks such as resource allocation, beamforming, power control, and even higher-layer protocols.
* **Application:**
  + Jointly training pilot signals and channel estimators.32
  + Jointly optimizing beamforming, power allocation, and RIS configuration.7
  + DDPG-E2E jointly optimizing transmitter and receiver, implicitly handling channel estimation.10
  + Optimizing communication timing, content, and mode in V2X based on Value of Information (VoI), integrating control and communication decisions.36 Reference 29 proposes joint optimization of IoT device offloading, STAR-RIS amplitude and phase shift coefficients, and power control. Reference 30 studies joint optimization of satellite beamforming, UAV positioning, power allocation, and rate splitting.
* **Potential:** Achieve system-level performance gains unattainable by optimizing individual components, leading to more integrated and intelligent communication systems. The trend of using RL for joint optimization and end-to-end learning 10 signals a potential paradigm shift where the traditional modular design of communication systems (separate channel estimation, equalization, decoding, etc.) might be replaced by more holistic, AI-driven designs that learn to perform these tasks implicitly and integratively. DDPG-E2E 10 has already demonstrated this by learning to communicate over unknown channels without an explicit estimation block. Extending this philosophy, RL could dynamically orchestrate various physical and MAC layer functions, leading to truly "intelligent" transceivers. This is a significant departure from current engineering practices.
* **Related Research:** 10 (DDPG-E2E), 32 (Joint pilot and estimator training), 34 (Joint beamforming and ADC threshold optimization), 36 (Joint communication and control optimization for CAVs).

However, there is a potential tension between the pursuit of more powerful and complex DRL models (e.g., for joint optimization or handling very complex states) and the need for lightweight, interpretable models. While sophisticated DRL architectures might achieve higher peak performance in specific complex tasks (e.g., 30 using TRPO, TD3, A3C for multi-parameter optimization), they can be harder to deploy 11, debug, and trust.11 The research community will need to find balances or develop hybrid approaches that offer both capability and practicality/interpretability.

Table 3 outlines the innovative frontiers and future research directions for RL in channel estimation.

**Table 3: Innovative Frontiers and Future Research Directions for RL in Channel Estimation**

| **Innovation Area** | **Specific RL Techniques/Concepts to Explore (e.g., MAML, Reptile, Offline Meta-RL; Parameter Transfer, Instance Transfer; Federated Averaging with RL, Multi-Agent RL; CNTK, Pruning/Quantizing DRL; Adversarial RL, Robust MDPs; Attention-based XAI, Surrogate Models; Multi-Objective RL, Hierarchical RL)** | **Potential Impact on Channel Estimation (e.g., Rapid adaptation to unknown channels, Reduced data requirements, Privacy & scalability, On-device deployment, Resilience to interference/attacks, Trustworthy AI, System-level performance gains)** | **Key Challenges to Address (e.g., Meta-task definition, Negative transfer, Communication overhead in FL, Performance retention in lightweight models, Defining appropriate threat models, Interpretability vs. performance trade-off, Complexity of joint optimization)** | **Related References** |
| --- | --- | --- | --- | --- |
| Meta-Reinforcement Learning (Meta-RL) | MAML, Reptile, Offline Meta-RL, Context-based Meta-RL with MoE | Rapid adaptation to unknown channels, Reduced training overhead for new environments, Improved generalization | Meta-task definition, Stability of few-shot learning, Computational complexity | 48 |
| Transfer Learning (TL) | Parameter transfer, Instance transfer, Feature representation transfer, Combined with DRL (DTL) | Reduced reliance on large target domain data, Accelerated learning, Improved performance in data-scarce scenarios | Negative transfer risk, Measuring relevance between source/target tasks, Effectiveness of knowledge transfer | 8 |
| Federated/Distributed RL | Federated Averaging with RL, Multi-Agent RL (MARL), Qmix, Heterogeneous FL (HFL) | Privacy preservation, Scalability, Reduced communication overhead, Support for on-device learning | Communication overhead, Heterogeneity management, Convergence guarantees, Incentive mechanism design | 4 |
| Lightweight DRL | Simplified NN architectures (Lightweight CNN/Transformer), Infinite-width NN (CNTK), Algorithmic optimization (e.g., sub-block partitioning), Model compression (pruning, quantization, knowledge distillation) | On-device deployment, Reduced latency, Improved feasibility for real-time applications | Maintaining high performance while reducing complexity, Integration of model compression with RL | 13 |
| Robust RL (for harsh environments) | RL algorithms for dynamic environments (DDPG), Robustness objectives/constraints, Combination with interference suppression (RSMA), Secure RL (adversarial training, robust reward design) | Reliable communication in non-stationary, high-interference, adversarial settings | Defining appropriate threat models, Robustness vs. performance trade-off, Generation and defense against adversarial samples | 1 |
| Explainable RL (XAI for CE) | XAI techniques for RL, Policy/value function visualization (e.g., attention maps), Inherently interpretable architectures, Learning world models | Enhanced system trust, Easier debugging & improvement, Insights into learned channel characteristics | Interpretability vs. performance trade-off, Difficulty in explaining complex RL models, Lack of standardized evaluation metrics | 11 |
| Joint CE & Cross-Layer Optimization | Joint pilot & estimator training, Joint beamforming/power/RIS configuration, E2E learning, Multi-objective RL, Hierarchical RL | System-level performance gains, More integrated and intelligent communication systems | Complexity of joint optimization problems, Multi-objective conflicts & trade-offs, Cross-layer information exchange | 7 |

## 7. Conclusion: The Transformative Impact of Reinforcement Learning on Channel Estimation

The introduction of Reinforcement Learning is profoundly changing the theory and practice of channel estimation in wireless communications. By endowing communication systems with unprecedented self-learning and adaptive capabilities, RL not only overcomes many limitations of traditional estimation algorithms in increasingly complex wireless environments but also lays a solid foundation for the construction of future intelligent communication systems.

Reviewing RL's contributions to channel estimation, its core value lies in effectively addressing dynamic, model-unknown, or model-complex channel scenarios. Whether optimizing symbol selection and resource allocation in MIMO and NOMA systems through Q-learning and its variants 15, or enabling intelligent decision-making in IRS/RIS configuration 1 and end-to-end communication 10 using advanced Actor-Critic methods like DDPG and PPO, RL has demonstrated immense potential in improving estimation accuracy, enhancing system robustness, and optimizing overall communication performance. Its preliminary explorations in time-varying channel tracking 12, efficient Massive MIMO estimation 13, and emerging fields like mmWave, IoT, and V2X 9 all signal the formation of a data-driven, more intelligent physical layer.

Looking ahead, the application of advanced learning paradigms such as meta-RL, transfer learning, and federated learning will further enhance the adaptability and efficiency of RL in channel estimation, enabling rapid response to new environments, utilization of prior knowledge, and protection of user privacy.48 The development of lightweight DRL algorithms will promote their deployment on resource-constrained devices, while research into robust RL and explainable AI aims to build more trustworthy and understandable agents.5 Crucially, RL-driven joint channel estimation and cross-layer optimization, such as end-to-end learning and the synergy of communication and control, foreshadow a potential shift in communication system design philosophy from traditional modular, separate optimization to more holistic, integrated AI-native designs.10

Although challenges remain in areas like computational complexity, sample efficiency, reward design, and security, ongoing research and innovation are continuously breaking through these bottlenecks. The ultimate goal of RL is to help build wireless networks that can autonomously perceive, understand, adapt to, and optimize their operational states. As these innovative directions deepen, RL is poised to play a central role in 6G and future communication systems, empowering truly intelligent, autonomous, and highly resilient wireless connectivity, thereby revolutionizing how we design and operate communication networks.

#### 引用的著作

1. Research on channel estimation based on joint perception and deep enhancement learning in complex communication scenarios - PeerJ, 访问时间为 五月 14, 2025， <https://peerj.com/articles/cs-2852.pdf>
2. Research on channel estimation based on joint perception and deep enhancement learning in complex communication scenarios - PeerJ, 访问时间为 五月 14, 2025， <https://peerj.com/articles/cs-2852/>
3. [2206.02165] A Survey on Deep Learning based Channel Estimation in Doubly Dispersive Environments - arXiv, 访问时间为 五月 14, 2025， <https://arxiv.org/abs/2206.02165>
4. CMC | Free Full-Text | Secure Channel Estimation Using Norm Estimation Model for 5G Next Generation Wireless Networks, 访问时间为 五月 14, 2025， <https://www.techscience.com/cmc/v82n1/59225/html>
5. Undermining Deep Learning Based Channel Estimation via Adversarial Wireless Signal Fabrication, 访问时间为 五月 14, 2025， <https://par.nsf.gov/servlets/purl/10339961>
6. Review on Channel Estimation for Reconfigurable Intelligent Surface Assisted Wireless Communication System - MDPI, 访问时间为 五月 14, 2025， <https://www.mdpi.com/2227-7390/11/14/3235>
7. A Survey on Reinforcement Learning for Reconfigurable Intelligent ..., 访问时间为 五月 14, 2025， <https://www.mdpi.com/1424-8220/23/5/2554>
8. deep reinforcement learning: an overview - arXiv, 访问时间为 五月 14, 2025， <https://arxiv.org/pdf/1701.07274>
9. Spectrum Sharing using Deep Reinforcement Learning in Vehicular Networks - arXiv, 访问时间为 五月 14, 2025， <https://arxiv.org/pdf/2410.12521>
10. DDPG-E2E: A Novel Policy Gradient Approach for End-to-End Communication Systems, 访问时间为 五月 14, 2025， <https://arxiv.org/html/2404.06257v2>
11. DDPG-E2E: A Novel Policy Gradient Approach for End-to-End Communication Systems, 访问时间为 五月 14, 2025， <https://www.aimodels.fyi/papers/arxiv/ddpg-e2e-novel-policy-gradient-approach-end>
12. Reinforcement Learning-Aided Channel Estimator in Time-Varying MIMO Systems - PMC, 访问时间为 五月 14, 2025， <https://pmc.ncbi.nlm.nih.gov/articles/PMC10304914/>
13. A Low-Complexity Algorithm for a Reinforcement Learning-Based Channel Estimator for MIMO Systems - PMC, 访问时间为 五月 14, 2025， <https://pmc.ncbi.nlm.nih.gov/articles/PMC9229451/>
14. Reinforcement Learning-Aided Channel Estimator in Time-Varying MIMO Systems - MDPI, 访问时间为 五月 14, 2025， <https://www.mdpi.com/1424-8220/23/12/5689>
15. A Study on the Impact of Integrating Reinforcement Learning for ..., 访问时间为 五月 14, 2025， <https://pmc.ncbi.nlm.nih.gov/articles/PMC9921540/>
16. Deep Reinforcement Learning for Optimizing Restricted Access Window in IEEE 802.11ah MAC Layer - MDPI, 访问时间为 五月 14, 2025， <https://www.mdpi.com/1424-8220/24/10/3031>
17. arxiv.org, 访问时间为 五月 14, 2025， <https://arxiv.org/abs/2101.10300>
18. Simplified Deep Reinforcement Learning Approach for Channel Prediction in Power Domain NOMA System - MDPI, 访问时间为 五月 14, 2025， <https://www.mdpi.com/1424-8220/23/21/9010>
19. arxiv.org, 访问时间为 五月 14, 2025， <https://arxiv.org/pdf/2309.12507>
20. Deep Reinforcement Learning for Dynamic Multichannel Access in Wireless Networks, 访问时间为 五月 14, 2025， <https://www.researchgate.net/publication/323302666_Deep_Reinforcement_Learning_for_Dynamic_Multichannel_Access_in_Wireless_Networks>
21. Energy-Efficient Irregular RIS-aided UAV-Assisted Optimization: A Deep Reinforcement Learning Approach - arXiv, 访问时间为 五月 14, 2025， <https://arxiv.org/html/2504.15031v1>
22. DDPG-E2E: A Novel Policy Gradient Approach for End-to-End Communication Systems, 访问时间为 五月 14, 2025， <https://www.researchgate.net/publication/385231114_DDPG-E2E_A_Novel_Policy_Gradient_Approach_for_End-to-End_Communication_Systems>
23. Nguyen HUYNH | Lecturer | PhD | University of Liverpool, Liverpool | UoL | Department of Electrical Engineering and Electronics | Research profile - ResearchGate, 访问时间为 五月 14, 2025， <https://www.researchgate.net/profile/Nguyen-Huynh-4>
24. 访问时间为 一月 1, 1970， httpshttps://arxiv.org/html/2404.06257v2
25. 访问时间为 一月 1, 1970， <https://arxiv.org/pdf/2404.06257>
26. A PPO-based Channel Hopping Sequence Framework for Time Slot Channel Hopping - ResearchGate, 访问时间为 五月 14, 2025， <https://www.researchgate.net/publication/384851434_A_PPO-based_Channel_Hopping_Sequence_Framework_for_Time_Slot_Channel_Hopping>
27. PPO Simply Explained for Beginners with Detailed Walkthrough (Proximal Policy Optimization) - YouTube, 访问时间为 五月 14, 2025， <https://www.youtube.com/watch?v=5VHLd9eCZ-w>
28. Using Deep Reinforcement Learning to Enhance Channel Sampling Patterns in Integrated Sensing and Communication - arXiv, 访问时间为 五月 14, 2025， <https://arxiv.org/html/2412.03157v1>
29. Proximal Policy Optimization for Energy-Efficient MEC Systems with STAR-RIS Assistance, 访问时间为 五月 14, 2025， <https://icoin.org/media?key=site/icoin2024/abs/A-3-3.pdf>
30. Energy Efficient RSMA-Based LEO Satellite Communications Assisted by UAV-Mounted BD-Active RIS: A DRL Approach - arXiv, 访问时间为 五月 14, 2025， <https://www.arxiv.org/pdf/2505.04148>
31. Deep Residual Learning for Channel Estimation in Intelligent Reflecting Surface-Assisted Multi-User Communications - arXiv, 访问时间为 五月 14, 2025， <https://arxiv.org/pdf/2009.01423>
32. An Attention-Aided Deep Learning Framework for Massive MIMO Channel Estimation - arXiv, 访问时间为 五月 14, 2025， <https://arxiv.org/pdf/2108.09430>
33. Deep Learning-Based Channel Estimation - arXiv, 访问时间为 五月 14, 2025， <https://arxiv.org/pdf/1810.05893>
34. www.arxiv.org, 访问时间为 五月 14, 2025， <https://www.arxiv.org/pdf/2504.18957>
35. Channel Estimation in RIS-Enabled mmWave Wireless Systems: A Variational Inference Approach - arXiv, 访问时间为 五月 14, 2025， <https://arxiv.org/pdf/2308.13616>
36. Learning Value of Information towards Joint Communication and Control in 6G V2X - arXiv, 访问时间为 五月 14, 2025， <http://www.arxiv.org/pdf/2505.06978>
37. Learning Value of Information towards Joint Communication and Control in 6G V2X - arXiv, 访问时间为 五月 14, 2025， <https://arxiv.org/pdf/2505.06978>
38. Reinforcement Learning-Based Resource Allocation Scheme of NR-V2X Sidelink for Joint Communication and Sensing - MDPI, 访问时间为 五月 14, 2025， <https://www.mdpi.com/1424-8220/25/2/302>
39. Energy-Efficient and Intelligent ISAC in V2X Networks with Spiking Neural Networks-Driven DRL - arXiv, 访问时间为 五月 14, 2025， <https://arxiv.org/html/2501.01038v1>
40. Energy-Efficient and Intelligent ISAC in V2X Networks with Spiking Neural Networks-Driven DRL - arXiv, 访问时间为 五月 14, 2025， <https://arxiv.org/pdf/2501.01038>
41. Deep Reinforcement Learning-Based Adaptive Modulation for Underwater Acoustic Communication with Outdated Channel State Information - MDPI, 访问时间为 五月 14, 2025， <https://www.mdpi.com/2072-4292/14/16/3947>
42. 访问时间为 一月 1, 1970， <https://www.mdpi.com/2072-4292/14/16/3947/pdf>
43. Robust Channel Estimation for Optical Wireless Communications Using Neural Network, 访问时间为 五月 14, 2025， <https://arxiv.org/html/2504.02134v1>
44. Robust Channel Estimation for Optical Wireless Communications Using Neural Network - arXiv, 访问时间为 五月 14, 2025， <https://arxiv.org/pdf/2504.02134>
45. Robust Underwater Acoustic Channel Estimation Method Based on Bias-Free Convolutional Neural Network - ResearchGate, 访问时间为 五月 14, 2025， <https://www.researchgate.net/publication/377280019_Robust_Underwater_Acoustic_Channel_Estimation_Method_Based_on_Bias-Free_Convolutional_Neural_Network>
46. Channel Estimation by Infinite Width Convolutional Networks M. Mallik and G. Villemaud are with INSA Lyon, Inria, CITI, UR3720, France. (email:mohammed.mallik@insa-lyon.fr, guillaume.villemaud@insa-lyon.fr).This work was supported by a French government grant managed by the - arXiv, 访问时间为 五月 14, 2025， <https://arxiv.org/html/2504.08660v1>
47. channel estimation by infinite width convolutional networks - arXiv, 访问时间为 五月 14, 2025， [https://arxiv.org/pdf/2504.08660?](https://arxiv.org/pdf/2504.08660)
48. Age and Power Minimization via Meta-Deep Reinforcement Learning in UAV Networks - arXiv, 访问时间为 五月 14, 2025， [https://arxiv.org/pdf/2501.14603?](https://arxiv.org/pdf/2501.14603)
49. Meta-Reinforcement Learning With Mixture of Experts for Generalizable Multi Access in Heterogeneous Wireless Networks - arXiv, 访问时间为 五月 14, 2025， <https://arxiv.org/html/2412.03850v1>
50. Offline Meta-Reinforcement Learning with Evolving Gradient Agreement - Unpaywall, 访问时间为 五月 14, 2025， <https://unpaywall.org/10.1109%2FIROS58592.2024.10802167>
51. Reinforcement Learning in Process Industries: Review and Perspective, 访问时间为 五月 14, 2025， <https://www.ieee-jas.net/article/doi/10.1109/JAS.2024.124227>
52. Meta-learning based time-varying channel estimation method, 访问时间为 五月 14, 2025， <https://www.sys-ele.com/EN/10.12305/j.issn.1001-506X.2023.06.32>
53. From Theory to Practice: Implementing Meta-Learning in 6G Wireless Infrastructure, 访问时间为 五月 14, 2025， <https://www.researchgate.net/publication/386523661_From_Theory_to_Practice_Implementing_Meta-Learning_in_6G_Wireless_Infrastructure>
54. arxiv.org, 访问时间为 五月 14, 2025， <http://arxiv.org/pdf/1912.12265>
55. Coalition Formation for Heterogeneous Federated Learning Enabled Channel Estimation in RIS-assisted Cell-free MIMO - arXiv, 访问时间为 五月 14, 2025， <https://arxiv.org/html/2502.05538v1>
56. Meta-learning Approaches for Smart Antenna Systems in 5G Networks Using Reinforcement Learning and Artificial Intelligence - ResearchGate, 访问时间为 五月 14, 2025， <https://www.researchgate.net/publication/389639231_Meta-learning_Approaches_for_Smart_Antenna_Systems_in_5G_Networks_Using_Reinforcement_Learning_and_Artificial_Intelligence>
57. UAV-Aided Cellular Communications with Deep Reinforcement Learning Against Jamming | Request PDF - ResearchGate, 访问时间为 五月 14, 2025， <https://www.researchgate.net/publication/343730113_UAV-Aided_Cellular_Communications_with_Deep_Reinforcement_Learning_Against_Jamming>
58. rate splitting for interference channels with deep reinforcement learning - Middle East Technical University, 访问时间为 五月 14, 2025， <https://avesis.metu.edu.tr/dosya?id=7d2388a3-edfa-4214-a2cc-363324c4432d>
59. Predicting Future Actions of Reinforcement Learning Agents - NIPS papers, 访问时间为 五月 14, 2025， <https://proceedings.neurips.cc/paper_files/paper/2024/file/5aea56eefab60e06f35016478e21aae6-Paper-Conference.pdf>