



Review

Proximal Policy Optimization in 5G, B5G, and 6G Communication Systems: A Systematic Review

Vijaya Kittu Manda ^{1,2,*} , Bhukya Madhu ³ and Theodore Tarnanidis ⁴ ¹ Department of Computer Science, BEST Innovation University, Gorantla 515231, Andhra Pradesh, India² Department of Interdisciplinary Research, PBMEIT, Visakhapatnam 530027, Andhra Pradesh, India³ Department of CSE, Malla Reddy (MR) Deemed to be University, Hyderabad 500100, Telangana, India; madhu0525@gmail.com⁴ Department of Business Administration, University of Macedonia, 54006 Thessaloniki, Central Macedonia, Greece; tarnanidis@uom.edu.gr

* Correspondence: vijaykittu@hotmail.com or 2023scsc017@bestiu.edu.in; Tel.: +91-98495-19188

Abstract

Fifth-generation (5G), Beyond 5G (B5G), and sixth-generation (6G) wireless networks, along with the Internet of Things (IoT), are core communication infrastructure in smart cities. Their increased deployments create high-dimensional optimization and resource management challenges. Consequently, researchers have increasingly explored the use of Artificial Intelligence (AI) models for optimizing networks. The Proximal Policy Optimization (PPO) is one such algorithm that optimizes networks. This Systematic Literature Review (SLR) follows the PRISMA 2020 protocol to review 76 studies published between 2023 and 2026 to synthesize recent PPO-based approaches to optimize communication systems. This study examines key PPO variants in major communication domains. It outlines the primary obstacles to real-world deployment and provides a cross-domain classification. According to this study, PPO provides continuous action spaces with good training stability for AI models. Its stable policy-learning capabilities make it suitable for next-generation communication systems. However, sim-to-real transfer, reward design, and multi-agent scalability are a few key challenges encountered. Future directions emphasize robust, deployable PPO frameworks for 6G, IoT, and internet architecture.

Keywords: intelligent communication systems; proximal policy optimization; deep reinforcement learning; multi-agent systems; wireless communication systems; network optimization; integrated sensing; network slicing; beamforming; satellite communications



Academic Editor: Jose I. Moreno Novella

Received: 16 May 2026

Revised: 24 June 2026

Accepted: 25 June 2026

Published: 27 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The rapid rollout of 5G networks is expected to more than double global mobile data traffic, reaching over 300 exabytes per month by 2030 [1]. At the same time, 5G networks will carry more than 80% of all mobile traffic, creating unprecedented demand for intelligent, adaptive communication management. Since real-time data transfer is crucial, technologies such as Artificial Intelligence (AI) and the Internet of Things (IoT) play a vital role [2]. IoT connectivity is a cornerstone of smart city infrastructure. Sustainable management of spectrum and energy resources directly affects urban efficiency and resilience. AI methods such as machine learning (ML) and deep learning (DL) [3] are increasingly used in wireless networks [4] and for spectrum management [5]. This digital transformation of network management can reduce energy consumption and infrastructure costs [6–8]. Recent PPO-based optimization studies reported measurable energy-efficiency

improvements, including gains of 15.8% over DDPG and 34.2% over fractional programming baselines in RIS-assisted MU-MISO systems [9]. Specifically, reinforcement learning offers a dynamic framework for these methods to adapt to shifting network environments in real time [10]. Despite the growing use of deep reinforcement learning (DRL) in communication systems, existing review studies primarily examine either general DRL methods or isolated communication tasks such as routing, beamforming, or resource allocation. No recent review systematically synthesizes how Proximal Policy Optimization (PPO) supports energy-efficient and communication optimization across multiple domains in 5G, IoT, smart-city, and B5G/6G contexts.

To address the challenges of AI implementation in communication systems, this review focuses on Proximal Policy Optimization (PPO). PPO has emerged as a popular deep reinforcement learning algorithm used for high-dimensional, continuous-control optimization in modern communication networks. PPO is particularly relevant because it combines training stability with efficient handling of continuous decision variables such as transmit power, beamforming weights, routing priorities, and computation offloading. [11]. Its training stability often outweighs sample efficiency when compared to other algorithms [12,13]. Further, PPO has a remarkable ability to handle continuous action spaces inherent to wireless network problems, making it one of the most suitable approaches for many communication optimization problems [14].

The review integrates perspectives from various fields. The review examines PPO applications across wireless communications, edge computing, sensing, and control systems. Several recent reviews have examined reinforcement learning in wireless communication systems, DRL for resource allocation [15], or AI-enabled 6G networking [16]. However, these reviews either (i) treat PPO only briefly among many DRL algorithms, (ii) focus on a single application domain such as RIS, MEC, or routing, or (iii) do not compare PPO variants across multiple communication domains. In contrast, this review specifically synthesizes PPO variants across multiple communication domains.

The review covers communication, sensing, computation, and control domains within 5G, B5G, and 6G systems. Global 6G subscriptions are forecasted to reach 180 million by 2031 [17]. The first commercial deployments are expected around 2030 in countries such as the United States, China, Japan, South Korea, GCC nations, and India [18]. 6G is expected to support data rates up to 95 Gbit/s and eventually approach 1 Tbit/s with microsecond-level latency [19,20]. Applications of urban communication systems involving the deployment of IoT and edge devices are expected to be among the fastest-growing segments of the 6G ecosystem [17,21]. In particular, algorithms such as PPO can benefit in various areas in the communication networks domain.

Table 1 demonstrates that existing review studies primarily focus either on broad, deep reinforcement learning techniques or on isolated communication domains, such as resource allocation, RIS optimization, MEC, routing, or UAV-assisted communication systems. In contrast, the present review specifically synthesizes PPO and its major variants across multiple communication domains while integrating sustainability considerations, 6G/B5G perspectives, and cross-domain system-level analysis within a unified taxonomy.

This review treats PPO applications as a single unified framework for intelligent communication systems and not as separate optimization problems. These systems operate across different 5G, B5G, and 6G environments and require collaboration across multiple layers spanning multiple domains. The analysis outlines persistent architectural trends within these systems, specifically in adaptive resource management, policy learning, multi-agent coordination, and sustainability-aware network orchestration. These trends frequently appear across several communication domains such as beamforming, MEC, satellite systems, and V2X networks. Identifying them as interconnected systems rather

than isolated, domain-specific solutions enables systems developers to conduct a more in-depth systems-level study. This approach reveals deployment issues and optimization trade-offs that limit real-time deployment.

Table 1. Comparison of Existing Review Articles and the Present Study.

Review	PPO Specific	Multi-Domain Coverage	Sustainability Analysis	PPO Variant Comparison	6G/B5G Focus	Cross-Domain Taxonomy	Key Limitation
Cheng et al. (2023)—DRL for Wireless Resource Allocation [15]	X	Partial	X	X	Partial	X	Primarily focused on resource allocation and broad DRL methods.
Jiao et al. (2024)—AI-Enabled 6G Networking Review [16]	X	✓	Partial	X	✓	X	PPO was discussed only briefly among multiple AI techniques
Musaddiq et al. (2023)—RL-Based Routing and Spectrum Management Survey [22]	X	X	X	X	Partial	X	Limited to routing and spectrum-access applications
Puspitasari and Lee (2023)—DRL for RIS and Beamforming Optimization Review [23]	X	X	Partial	X	✓	X	Domain-specific focus on RIS and beamforming
Ismail et al. (2025)—MEC and Computation Offloading Using DRL Survey [24]	X	X	X	X	Partial	X	Focused exclusively on MEC and offloading systems
Hady et al. (2025)—Multi-Agent Reinforcement Learning for Wireless Networks Review [25]	Partial	Partial	X	Partial	✓	X	Limited comparison of PPO variants across domains
Amodu et al. (2025)—RL-Based UAV Communication Systems Survey [26]	X	X	Partial	X	Partial	X	Restricted to UAV-assisted communication networks
PPO in Intelligent Communication Systems (This Review)	✓	✓	✓	✓	✓	✓	Provides integrated cross-domain synthesis of PPO architectures and deployment challenges

Novel Contributions of This Review

Although RL in wireless communications has attracted considerable scholarly interest, most current studies have focused on general categories of deep RL algorithms. In particular, the use of PPO for communication system optimization is understudied. This review goes beyond a mere cataloging of PPO applications. It presents a cross-domain taxonomy that links the communication task, the PPO variant used, the implemented reward structure, the targeted sustainability parameter, and the encountered deployment challenge. This analysis juxtaposes PPO, MAPPO, IPPO, DPPO, hybrid PPO-DQN, and PPO-Transformer designs across nine communication domains. Challenges such as the sim-to-real transfer gap, the underreported energy cost of training large PPO models, and the multi-agent scalability ceiling in ultra-dense 6G environments are explicitly identified and positioned as frontier problems that the next wave of research must confront.

2. Technical Background of PPO for Communication Systems

PPO provides a robust framework for autonomous decision-making in highly dynamic environments with continuous action spaces (e.g., power levels, beamforming vectors). PPO offers several advantages over many alternative deep reinforcement learning methods. For example, Trust Region Policy Optimization (TRPO) requires solving a constrained optimization problem via conjugate gradients. PPO achieves similar stability guarantees

using only first-order optimization and is far simpler to implement. It provides stability in training, sample efficiency, and native support for continuous action spaces. Further, it can handle high-dimensional state spaces, which are common in dynamic network environments. Owing to these reasons, PPO is used as a foundational DRL algorithm for communication networks [27].

2.1. Applications of PPO in Communication Systems

Communication systems present unique challenges such as rapidly changing channel conditions, interference, and stringent quality of service requirements. Table 2 presents the key communication features and how PPO is particularly well-suited for these tasks.

Several highly cited communication-specific papers (e.g., the PPO-ISAC-THz paper with 84 citations) confirm that PPO has emerged as a widely adopted approach for complex mixed-integer and continuous optimization problems across next-generation wireless systems [28]. PPO applications span diverse fields, including satellite power management, network slicing, and drone trajectory planning. This widespread adoption demonstrates the algorithm’s flexibility as a primary framework for developing intelligent communication systems. Figure 1 illustrates the system-level operational framework of PPO in intelligent communication systems. The framework highlights how PPO-based agents continuously interact with dynamic communication environments by collecting real-time states, optimizing policies, taking adaptive control actions, and improving performance with feedback, thereby supporting adaptive resource allocation and energy-aware communication optimization in 5G/B5G/6G infrastructures.

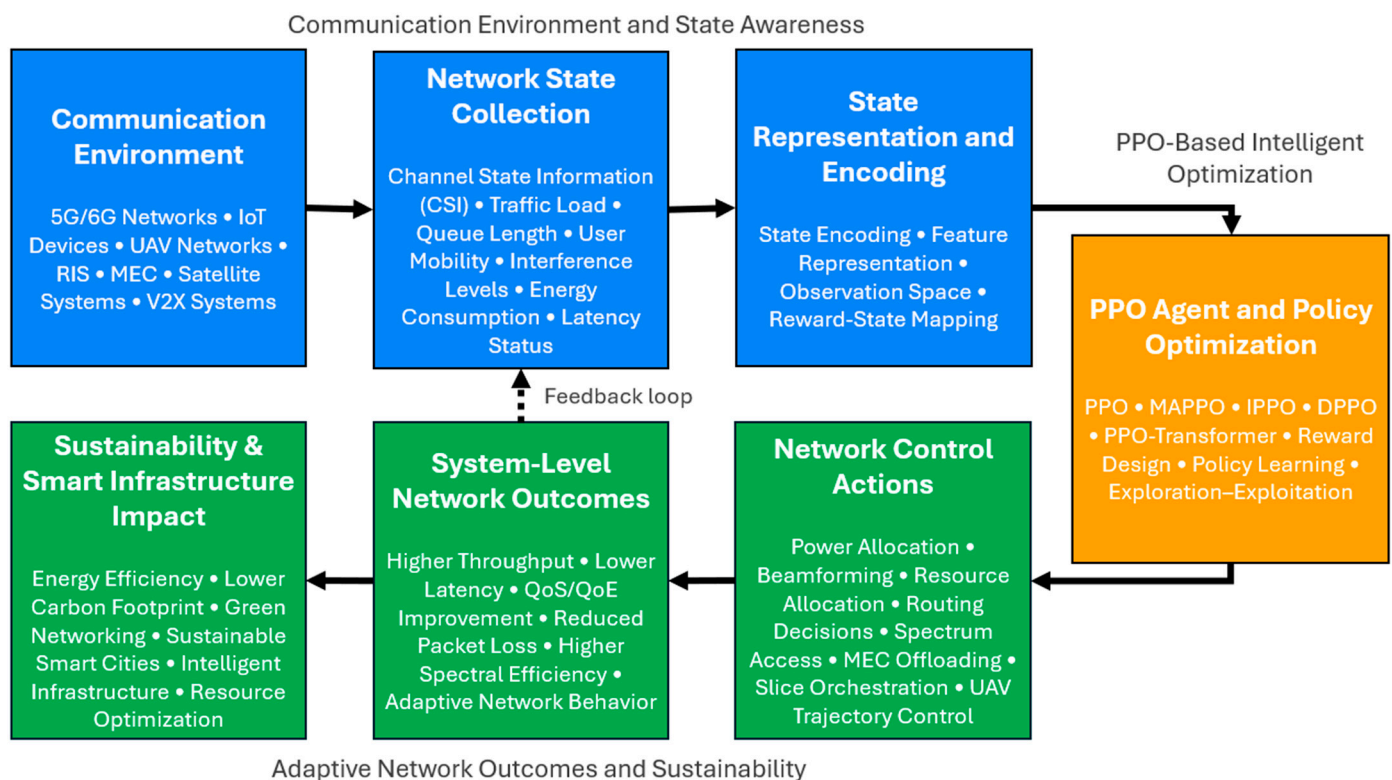


Figure 1. System-Level PPO Framework for Intelligent Communication Systems.

Table 2. Benefits of using PPO in Communications.

Feature	Benefit in Communications
Clipped Objective Function	Prevents large, destructive policy updates, ensuring stable learning even in volatile wireless environments [12,13].
Continuous Action Spaces	Unlike DQN, PPO can directly optimize continuous variables, such as transmit power and phase shifts [11,12].
Sample Efficiency	Allows for multiple epochs of updates on the same batch of data, reducing the need for massive real-time datasets [12,13].
Ease of Tuning	Generally, requires less hyperparameter tuning than algorithms like DDPG or SAC, making it more practical for deployment [13].
Energy Efficiency	PPO-based optimization improved energy efficiency by 15.8% over DDPG in RIS-assisted MU-MISO systems [9].

2.2. Alternatives to PPO

Several alternative deep reinforcement learning algorithms were developed for use in communication networks [3]. Alongside PPO, other popular ones include Soft Actor-Critic (SAC), Deep Deterministic Policy Gradient (DDPG), Twin Delayed DDPG (TD3), Advantage Actor-Critic (A3C), and Trust Region Policy Optimization (TRPO) [24]. This review specifically focuses on PPO because of its widespread adoption in recent 5G, B5G, and 6G communication studies. SAC generally provides higher sample efficiency through off-policy learning and entropy-based exploration, while TD3 and DDPG can achieve strong performance in continuous-control environments. However, these methods often require more careful hyperparameter tuning. Further, they exhibit greater training instability in highly dynamic wireless environments. PPO offers a practical balance between stability, implementation simplicity, computational efficiency, and continuous-action optimization. These features explain its growing popularity across communication domains, as discussed in Section 4. Therefore, this review focuses on PPO and its major variants (MAPPO, IPPO, DPPO, PPO-Transformer, and hybrid PPO architectures). We do acknowledge the importance of alternative DRL approaches as complementary research directions. Table 3 gives a brief comparison of various algorithms.

Table 3. Comparison of Proximal Policy Optimization (PPO) with Alternative Deep Reinforcement Learning Algorithms for Communication Systems.

Algorithm	Key Strength	Limitation	Typical Communication Use
SAC	High sample efficiency, entropy-driven exploration, and strong performance in continuous control	More hyperparameter tuning, higher implementation complexity	Power control, beamforming, dynamic spectrum management, UAV communications
TD3	Mitigates overestimation bias, improved stability over DDPG	Sensitive to parameter selection, increased architectural complexity	Resource allocation, RIS optimization, continuous control problems
DDPG	Efficient continuous-action optimization, lower interaction cost	Training instability, exploration challenges, prone to local optima	Power allocation, computation offloading, and beamforming
PPO	Stable training, easy implementation, strong continuous-action support, robust convergence	Lower sample efficiency than off-policy methods	Resource allocation, beamforming, network slicing, MEC, V2X, satellite communications
TRPO	Strong theoretical guarantees and monotonic policy improvement	Computationally expensive, difficult implementation due to constrained optimization	Early wireless resource management and continuous-control optimization
A3C	Parallel learning, reduced training time, simple architecture	Less stable than PPO, lower sample efficiency	Routing, congestion control, adaptive network management

3. Materials and Methods

The review methodology reports the databases, search strings, inclusion criteria, and PRISMA screening stages to ensure reproducibility.

3.1. Review Design and Research Questions

The study uses the Systematic Literature Review (SLR) approach, following the PRISMA 2020 guidelines [29]. It comprises the standard PRISMA checklist for core reporting requirements. The checklist is made available as a supplement to this document. The review protocol was not prospectively registered.

Objective of the Review: The main objective of this review is to synthesize recent applications of the Proximal Policy Optimization (PPO) algorithm across modern communication systems and networks. A secondary objective is to identify key limitations and research directions that will help shape the future research requirements for building next-generation sustainable networks.

Research Questions (RQs): Based on the research objectives, three research questions are framed to identify the current state of PPO applications in communication systems and to identify emerging future directions.

RQ1: *What are the major application domains and recurring design patterns of PPO-based approaches in modern communication systems?*

RQ2: *What key limitations and emerging future research directions are shaping the development of sustainable PPO-based communication frameworks for 5G, B5G, and 6G networks?*

RQ3: *How do PPO-based communication frameworks contribute to sustainable digital infrastructure, smart cities, and IoT system design?*

3.2. Search Strategy

Search Databases Used: Scopus (n = 108) and Web of Science (n = 475) are the primary databases used for the literature review search. Web of Science yielded a larger number of records because the Topic search covered multiple indexed collections and broader metadata fields than the Scopus search. Google Scholar (n = 57) is used to selectively identify recent high-quality publications from reputed publishers that are not yet indexed in the primary databases. Greater emphasis was placed on studies published through IEEE Xplore, Springer, MDPI, Wiley, Association for Computing Machinery (ACM), and ScienceDirect. To ensure consistency across databases, the same search keywords and Boolean operators were used throughout the literature search. In Scopus, the search was performed using the fields “Article Title, Abstract, and Keywords”. In Web of Science, the search was conducted using “Title, Abstract, and Author Keywords”. Google Scholar was also used as a supplementary search source to identify potentially relevant studies that may not have been indexed in the primary databases. Because Google Scholar does not offer the same structured field-selection capabilities as Scopus and Web of Science, searches were conducted using the same keyword combinations, and results were manually screened for relevance. Records retrieved from all databases were subsequently screened for duplicates and assessed against predefined inclusion and exclusion criteria.

Search Keywords and Strings: The searches were performed using the following Boolean keyword combinations:

(“Proximal Policy Optimization” OR “PPO”) AND (“wireless communication” OR “communication systems” OR “5G” OR “B5G” OR “6G”)

(“MAPPO” OR “IPPO”) AND (“resource allocation” OR “routing” OR “dynamic spectrum access” OR “beamforming”)

(“PPO”) AND (“RIS” OR “MEC” OR “ISAC” OR “UAV” OR “V2X” OR “satellite communications” OR “edge computing”)

("deep reinforcement learning") AND ("communication networks") AND ("PPO") ("PPO") AND ("network slicing" OR "computation offloading" OR "cognitive radio") ("PPO" OR "deep reinforcement learning") AND ("energy efficiency" OR "smart city" OR "IoT" OR "sustainable network" OR "green communication")

Search Timeframe & Search Date: The review considered studies published between 2023 and 2026 to capture the most recent developments in PPO-based communication systems. This timeframe is essential for a sustainability-focused review, given that many "green" technologies, such as RIS and ISAC, only reached maturity in these years. The literature search was conducted on 22 June 2026.

3.3. Study Selection Criteria

Inclusion Criteria: Selected studies (especially those retrieved from Google Scholar) that explicitly focused on PPO algorithms and PPO variants in the communications domain are included to supplement the previously obtained search results from the primary databases. We prioritized peer-reviewed journal articles and conference proceedings published between 2023 and 2026. Priority was given to publications that demonstrated empirical performance benchmarks, novel architectural designs, or significant conceptual contributions relevant to PPO-enabled communication systems.

Exclusion Criteria: Duplicate records common across multiple database searches are removed prior to screening. Similarly, publications that are not predominantly related to communications domains or that lack DRL methodologies are excluded. Works lacking sufficient methodological detail or technical detail regarding algorithmic implementation, or not providing sufficient empirical analysis or simulation result analysis, are omitted entirely. The exclusion criteria helped remove many publications and ensured the study focused on high-quality works only.

Language Restrictions: To ensure consistent interpretation of technical concepts, only publications written in the English language were included. The restriction helped maintain uniformity during the data extraction and synthesis phases of the review.

Document Types Considered: The search included archival journal articles, conference proceedings, and early-access articles from major databases. Book chapters are considered, but with lower priority, only to support theoretical or foundational aspects. Editorials and unpublished, non-peer-reviewed literature (such as preprints on arXiv and TechRxiv) are excluded to maintain high scientific rigor.

3.4. Screening Process (PRISMA Core)

Initial Retrieval Count: The initial search strategy yielded 640 records across the selected digital libraries and adequately covered recent developments in PPO.

Duplicate Removal Process: Spreadsheet software (Microsoft Excel (v2605, Microsoft 365)) helped quickly identify and remove 180 duplicate entries from the combined search results. Though tedious, manual verification is then conducted to ensure that no unique studies are accidentally discarded during cleaning.

Title and Abstract Screening: Reviewers screened the titles and abstracts individually to assess relevance to our research questions. Cohen's kappa coefficient helped in statistically determining inter-rater reliability when agreement is reached. The title and abstract screening stage yielded a kappa value of 0.82, indicating strong agreement. Disagreements that arise are resolved through discussion and, where necessary, consultation with an additional reviewer. As many as 300 papers were excluded for lacking clear connections to communication systems or for failing to address optimization tasks at this stage.

Full-Text Eligibility Assessment: The remaining 160 articles underwent a rigorous full-text assessment to verify compliance with all inclusion criteria. Interrater agreement

for the full-text eligibility stage was also evaluated using Cohen's kappa, yielding a value of 0.86. The remaining disagreements were resolved through consensus. Then, 84 were excluded for non-compliance with the inclusion criteria. We evaluated methodological soundness and clarity of reported experimental results during this detailed examination phase. Of the 84 excluded studies, 38 were outside the (strict) scope of PPO-enabled communication systems, 26 lacked sufficient methodological detail, and 20 did not provide adequate empirical evaluation.

Final Study Count: The final selection yielded 76 qualified studies for the comprehensive qualitative synthesis analysis.

3.5. PRISMA Flow Diagram

The PRISMA flow diagram in Figure 2 visually explains the progression of the review from identification through screening to eligibility and inclusion. It follows the PRISMA 2020 study selection workflow. It presents the identification, screening, eligibility, and inclusion stages in a customized visual format that improves readability while preserving all required PRISMA reporting elements. It brings transparency into the study selection process and the reasons for excluding specific records. The initial retrieval count consisted of 108 records from Scopus and 475 records from Web of Science. A supplementary Google Scholar search yielded an additional 57 records. The combined set resulted in 640 records. Entries that are duplicates in the databases are identified using Microsoft Excel and removed. It eliminated 180 records, leaving 460 unique studies for screening. Title and abstract screening then excluded 300 papers because they lacked a clear connection to communication systems, PPO, or deep reinforcement learning. Consequently, 160 articles remained for full-text eligibility assessment. During this stage, 84 studies were excluded. Specifically, 38 studies were outside the review scope, 26 lacked sufficient methodological detail, and 20 did not provide adequate empirical evaluation. The final curated dataset consisted of 76 peer-reviewed studies used for qualitative synthesis and thematic analysis.

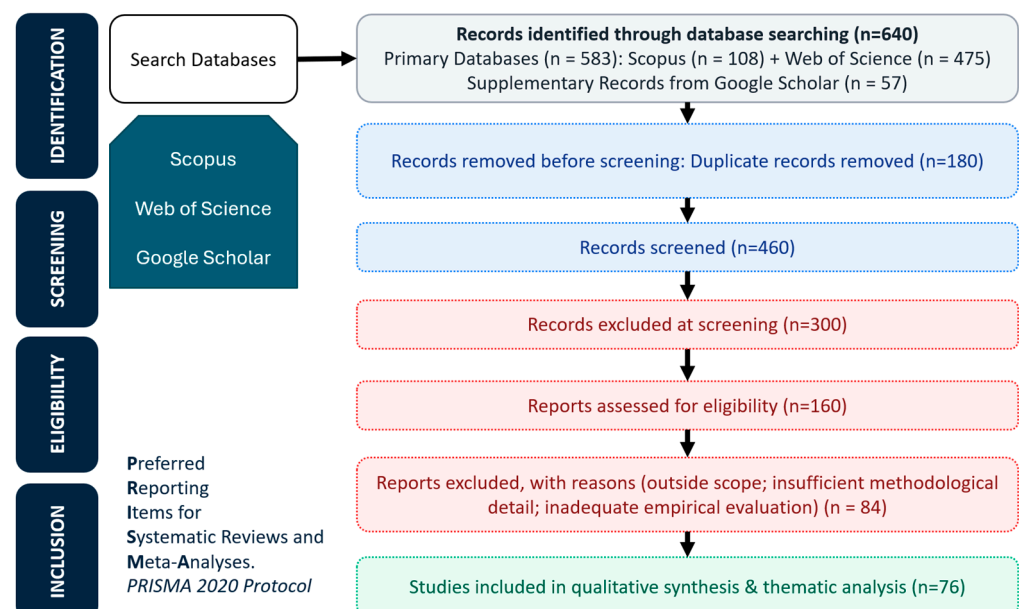


Figure 2. PRISMA Flow Diagram.

3.6. Data Extraction and Analysis

Data Extraction Criteria: The review synthesizes performance trends and identifies recurring design patterns in PPO-based communication frameworks. From each study, specific algorithmic details, such as the PPO variant and the associated hyperparameters,

are collected. Data extraction was conducted independently by the reviewers using a structured extraction framework. Disagreements, if any, between the authors were resolved through discussion and consensus. Structured data extraction is used to collect data uniformly across studies. Appendix A provides the complete list of studies in which PPO, MAPPO, IPPO, DPPO, or another named PPO variant was demonstrably implemented or analyzed as the primary or comparator algorithm. This is the qualitative-synthesis set used for the quantitative and architectural comparisons in Sections 4 and 5. A separate set of contextual references, including survey and review articles on broader DRL/MARL studies that do not isolate PPO as the evaluated algorithm, and perspective papers on sustainability, governance, or deployment, is cited throughout the narrative text. These are added to motivate the choice of DRL/PPO and to help position it against alternative RL approaches. It also supports claims about system architecture and deployment context. These contextual references are listed in the main reference list. However, they are excluded from Appendix A, since they do not strictly meet the PPO-implementation criterion used for the structured comparison.

Classification Framework/Taxonomy: The selected studies are divided into major application domains in contemporary wireless communication systems. The main groups are resource allocation, intelligent routing, beamforming, and multi-access edge computing.

Quality Assessment and Risk of Bias: The included studies were qualitatively assessed based on methodological clarity, reproducibility, empirical evaluation, and completeness of the reported experimental settings. Because the reviewed studies primarily consisted of engineering simulation studies rather than clinical trials, formal medical risk-of-bias instruments were not directly applicable. Potential reporting bias was minimized through multi-database searching and predefined inclusion criteria.

Certainty/Confidence Statement: Due to substantial heterogeneity across datasets, simulation environments, evaluation metrics, and communication scenarios, a formal certainty-of-evidence assessment was not performed.

Synthesis Approach: The retained studies are synthesized using thematic analysis. The thematic coding process used four predefined categories derived from the research questions:

- (i) Communication application domain;
- (ii) PPO variant and architectural design;
- (iii) Primary system benefit;
- (iv) Recurring technical limitation or future research direction.

Additional subthemes were identified inductively during full-text review. In this, publications are grouped by communication domain, PPO variant, sustainability contribution, and recurring technical challenges. The insights are derived through comparative analysis of empirical performance benchmarks across the selected studies. The review further identified recurring design patterns, including hybrid PPO architectures and centralized training with decentralized execution. Studies were coded independently into these categories, after which recurrent patterns were grouped into the thematic domains presented in Section 4.

Data Availability Statement: No new datasets were generated during the current study. The analyzed studies are available through the cited references.

4. Results and Discussion

Studies confirmed the feasibility of Multi-agent PPO (MAPPO) in real-world 5G systems, proving its effectiveness when combined with domain-aware observation design and sustainable network management goals [30,31]. Spectrum scarcity is worsening due to surging wireless traffic. In the upcoming years, it is anticipated that the total amount of mobile data traffic worldwide, excluding Fixed Wireless Access (FWA), will increase dramatically. It is projected to increase by approximately 2.3 times, reaching nearly

280 exabytes per month by 2030 [32]. This rapid growth is driven by rising demand for data-intensive applications such as video streaming, IoT services, and immersive communication technologies. Hence, there is an urgent need to address Dynamic Spectrum Access (DSA) in Cognitive Radio Networks (CRNs) to enhance spectrum efficiency. Most prior DSA approaches assume homogeneous Secondary Users (SUs). Deep reinforcement learning can help in jointly optimizing sensing and access policies under partial observations [33]. These approaches learn coordinated strategies without requiring prior knowledge of channel state statistics [34].

Fully decentralized multi-agent reinforcement learning frameworks effectively capture the distributed nature of participating stations [35]. One-bit gradient quantization reduces communication costs by more than an order of magnitude. These methods achieve performance comparable to centralized training frameworks while preserving complete decentralization. Further, prior DSA approaches ignored real-world heterogeneity. A Multi-agent PPO (MAPPO) can handle agent heterogeneity without parameter sharing. Further improvements made to the MAPPO helped achieve higher average reward (throughput) and provide efficient, stable, and fair spectrum access [36].

Researchers compared various PPO variants for use in the communications domain. For example, in a study, Independent PPO (IPPO), MAPPO, and a MAPPO + Transformer variant (called DET-FAIR-WINGS) are compared [30]. The study demonstrated that MAPPO-based approaches outperform IPPO in terms of convergence speed, packet loss, and latency. Several studies went beyond routing and spectrum access. PPO-based methods are applied to congestion control mechanisms such as Explicit Congestion Notification (ECN) [37]. ECN is a mechanism in computer networking that allows routers or switches to signal impending congestion to end hosts without dropping packets. The idea is to improve network performance by enabling early congestion signaling. Doing this can reduce the need for packet loss as a congestion-control signal, thereby helping in dynamic data-center communication environments. The existing static ECN thresholds are ineffective and inadequate for handling many-to-one (in-cast) traffic and heterogeneous flow types (mice vs. elephant flows). Combining offline and online training to adapt is found to benefit in dynamic network conditions. This was made possible with a multi-agent Independent Proximal Policy Optimization (IPPO) algorithm [37].

Studies extended PPO beyond its conventional use by embedding spatiotemporal reasoning directly into the policy network. This enables more intelligent, adaptive, and QoS-aware routing decisions in highly dynamic tactical networks [27]. Furthermore, such enhancements enable autonomous learning to adapt to dynamic traffic conditions. It also supports finer-grained routing control, thereby effectively reducing congestion. PPO optimizes probabilistic routing decisions across multiple nodes, thereby improving the routing approach used in communication devices. For example, it can integrate well with the actor-critic architecture, using Graph Attention Networks (GAT) to handle the complexities of network topologies [38]. It can have autonomous learning capabilities to adapt to dynamic traffic conditions and support finer-grained routing control. Eventually, it can effectively reduce congestion and resource wastage [39].

Recent work extends PPO to large-scale, high-mobility vehicular networks [40,41]. These environments require joint optimization of sensing, communication, and computing under dynamic conditions—similarly, cross-layer multi-agent PPO frameworks used dynamic Time Division Duplexing for enhanced resource allocation [42]. Centralized training with distributed execution optimizes uplink and downlink resource allocation while incorporating conditional activation. Simulation results demonstrate an 18% increase in throughput and a 16% reduction in interference.

Recent studies extend this approach by integrating sensing and communication within a unified learning framework [43]. Energy-efficient DRL-based beamforming has been explored for V2X scenarios under dynamic and uncertain channel conditions. PPO-driven optimization significantly reduces the average AoI compared to existing baseline approaches, thereby effectively balancing sustainable energy harvesting with data freshness [44,45].

Similarly, challenging dynamic environments are found in maritime wireless networks [46]. Devices on navy ships, in shipping, and on drones keep moving, and their channels change quickly [33,47]. These platforms also operate under strict power and bandwidth constraints. Therefore, traditional methods for assigning these resources will not be effective in such unpredictable environments. Adaptive algorithms adjust to dynamic channel conditions to improve reliability and throughput across mobile nodes. These characteristics make maritime networks a natural application domain for adaptive RL algorithms such as PPO [48]. Adaptive PPO-based resource allocation can reduce unnecessary power transmission and improve bandwidth utilization in maritime communication systems [49]. Meta-learning techniques can improve PPO adaptation in dynamic sea conditions. Domain knowledge can also help PPO perform better with limited data. Table 4 presents a taxonomy of PPO applications in communication systems based on the communication domain, PPO variant, and the corresponding optimization mechanism or outcome. Additional discussions regarding optimization objectives, sustainability considerations, and deployment challenges are provided in the subsequent sections.

Table 4. Taxonomy of PPO-Based Applications in Intelligent Communication Systems.

Communication Domain	Typical PPO Variant	Key Mechanism/Outcome
Resource Allocation	PPO, MAPPO, PPO-Transformer	Reduces unnecessary power transmission and enables greener network operation [35,50].
Beamforming & RIS	MAPPO, IPPO	Hardware-efficient signal enhancement lowers overall energy requirements [9,28,51].
MEC & Offloading	PPO, DPPO	Balances local energy use with transmission latency for improved trade-offs [52–54].
Vehicular Edge Computing & Task Offloading	PPO, MAPPO	Joint optimization of computation offloading, resource allocation, latency reduction, and edge-resource utilization
Integrated Sensing (ISAC)	PPO, MAPPO, PPO-DQN	Reduces the need for dedicated sensing hardware in smart city infrastructure [55–57].
UAV Communications	PPO, Dual-Clip MAPPO	Energy management extends flight time and balances secrecy with efficiency [51,58].
Network Slicing	PPO, MAPPO	Enables adaptive multi-service resource orchestration [59–61].
Satellite Communications	PPO, PPO-DQN Hybrid	Reduces the energy cost of providing connectivity to remote regions [62–64].
V2X & Cognitive Radio	PPO, MAPPO	Reduces idle time, device battery drain, and overall transportation congestion [40,65,66].

4.1. Resource Allocation and Management

Resource allocation remains a central optimization challenge in next-generation networks [8]. Dynamic resource allocation is perhaps the most prominent use case with several active application areas in wireless systems. Addressing scalability challenges in dense 5G network planning, recent investigations introduce a hierarchical multi-agent PPO framework integrated with representation learning [67]. This architecture effectively decouples base station deployment from antenna parameter tuning while leveraging UNet mechanisms to process large-scale spatial information. Empirical evaluations demonstrate

significant gains in coverage rates and average throughput compared to conventional multi-agent baselines.

Several studies reported PPO outperforming heuristic algorithms across a range of user loads and channel conditions when jointly optimizing transmit power, bandwidth, and scheduling in heterogeneous networks (HetNets) [50,68]. In Open Radio Access Networks (O-RAN), PPO (on-policy) has been compared with ACER (off-policy) for QoS-sensitive resource allocation, with PPO demonstrating a favorable balance between energy consumption and user latency [69]. O-RAN uses a disaggregated architecture. In this setup, different xApps manage various RAN functions, such as power control and resource block allocation. These xApps need to coordinate their decisions without causing interference. A Digital Twin-enabled HAPPO framework successfully synchronizes them without requiring direct communication between the xApps [70]. The concepts can be extended this to energy sustainability. For example, a hierarchical rApp-xApp PPO framework has demonstrated that jointly optimizing radio unit activation schedules and user association weights [71]. Such system substantially reduces O-RAN operational power consumption making it ideal for 6G deployments. In adversarial scenarios, researchers proposed a new jamming estimator that uses deep hierarchical PPO and is integrated into the MEC-O-RAN framework [72]. This estimator models the ON-OFF jammer's behavior as a partially observable Markov process. It predicts which time slots will be jammed. This allows a Transformer-based task scheduler to work only on safe slots. As a result, the cumulative prediction error is only 13 time slots per 100, compared to 48 for standard DDQN. It also reduces task execution time from 1260 s to 316 s for 50 active users. By using self-attention to map user sequences to resource sequences, a Transformer-based PPO framework in multiuser cellular systems provides greater spectrum efficiency and user fairness [73]. In addition to these approaches, current research has investigated hybrid reinforcement learning strategies for handling multi-objective optimization issues. A hybrid reinforcement learning framework that combines PPO and deep Q-networks has been presented to increase exploration efficiency and policy stability in dynamic control contexts [74]. These techniques allow simultaneous optimization of conflicting goals, such as energy sustainability, system stability, and performance efficiency, by using multi-objective reward structures [75]. These results demonstrate the growing importance of incentive design and hybrid learning architectures for solving complex, practical optimization problems in control and communication systems.

Effective coordination between dispersed stations is enabled by quantized gradient updates that do not require exchanging local data [35]. Independent PPO successfully operates within both decentralized training and decentralized execution paradigms. Communication overhead is reduced substantially during agent training while maintaining competitive throughput levels. In UAV base station networks, a hierarchical PPO architecture decomposes a complex multi-channel allocation problem from exponential to linear complexity, achieving superior throughput and convergence.

Adaptive Bitrate (ABR) Video Streaming is a key example of QoE-aware resource management [76]. Providing high Quality of Experience (QoE) for video streaming over dynamic 5G networks is a challenging sequential decision problem well-suited to PPO. PPO-ABR is a system that addresses the lag between behavior and target policies in Asynchronous Advantage Actor-Critic (A3C) methods by using PPO's on-policy multi-agent DRL framework [77]. A related approach combines behavior cloning with PPO to accelerate convergence while maintaining stable bitrate selection under fluctuating network conditions. The method demonstrates improved QoE compared to benchmark schemes by using PPO's controlled policy update mechanism to adapt reliably to 5G network dynamics [78]. By maximizing sample efficiency through clipped probability ratios across

multiple epochs of minibatch updates, PPO-ABR outperforms state-of-the-art ABR methods across different QoE metrics on real network traces.

4.2. Adaptive Sensing and Beamforming Systems

Reconfigurable Intelligent Surfaces (RIS) technology reduces active transmission power requirements, making it a hardware-efficient and energy-sustainable approach to signal enhancement in dense urban environments. Hence, RIS hardware-efficient signal enhancement becomes a sustainable alternative to power-hungry active transmitters. PPO can continuously refine beamforming policies as channel conditions evolve [79]. The clipped-policy update mechanism helps prevent unstable beamforming decisions during training [80,81]. So, PPO is attractive for real-time beam management in highly dynamic wireless networks. Recent research explored PPO for RIS, in which the agent optimizes the surface's reflection coefficients to enhance signal quality. Further advancements have extended metasurface technology to stacked intelligent surfaces, enabling the effective use of enhanced PPO algorithms [51]. This approach effectively manages secure mobile communications. It helps in jointly optimizing power and phase shifts in dynamic environments. These methods achieve significant improvements in achievable secrecy rates compared to conventional DRL baselines.

PPO has become a leading solver for non-convex beamforming and RIS phase-shift optimization problems [51]. In an IRS-aided Integrated Sensing and Communications (ISAC) system in the THz band, a primal-dual PPO framework jointly optimizes continuous transmit beamforming (actor) and discrete phase-shift design (critic), with a distributed PPO (DPPO) variant for multiuser MIMO scenarios [82]. For RIS-assisted SWIPT with Rate Splitting Multiple Access (RSMA), a PPO-based approach simultaneously optimizes beamforming vectors, power splitting ratios, and discrete phase shifts. Simultaneous Wireless Information and Power Transfer (SWIPT) enables devices to receive both energy and data from the same wireless signal. PPO-optimization helped achieve energy efficiency close to that of successive convex approximation (SCA) methods while requiring substantially less runtime in time-varying channels [28]. Similarly, in an RIS-assisted MU-MISO system, the PPO algorithm outperforms fractional programming (FP) by 34.2% and DDPG by 15.8% in energy efficiency. PPO also proves effective for statistical CSI-based beamforming, significantly reducing channel estimation overhead compared to instantaneous CSI methods [9]. Such reductions in estimation overhead and transmission power directly lower the carbon footprint of dense urban network deployments.

4.3. Communication–Computation Co-Design

For IoT and vehicular networks, PPO can help decide whether a task should be processed locally or offloaded to an edge server [83]. MEC enables devices to offload compute-intensive tasks to nearby edge servers, recognizing that computation offloading is a central resource-scheduling problem [24]. Similarly, blockchain-enabled digital twin architectures have been explored to improve interoperability, trust management, and system coordination in distributed IoT and edge-computing environments [84]. Similar integrations are possible in edge intelligence and digital twin frameworks for IoV systems [85]. In such systems, RL enables reliable, energy-efficient coordination between communication and computing resources. Within these sustainable computing frameworks, PPO-based multi-agent reinforcement learning effectively solves resource-allocation problems while balancing delay, energy constraints, and energy-aware operations. These choices emphasize PPO's role in coordinating communication channels and allocating processing resources to achieve the best possible system performance. Recent work extends this paradigm by simultaneously optimizing inference execution and communication decisions in dispersed

edge settings. To jointly optimize DNN partitioning, uplink channel allocation, and transmission power regulation under shared wireless restrictions, a MAPPO framework has been suggested [52]. By adopting centralized training with decentralized execution and a multi-head critic design, the approach improves coordination among edge devices while maintaining scalability in large hybrid action spaces [86]. Experimental results demonstrate improved trade-offs between latency and energy. It highlights the effectiveness of PPO in tightly coupled communication–computation optimization scenarios within IoT edge systems [39,53].

The system adjusts to the current network load and server availability while balancing the trade-off between transmission latency and local energy consumption. Task deadlines, time-varying channels, and heterogeneous devices are common in MEC environments. It makes computation-offloading decisions challenging optimization problems [24]. Further, these situations are dynamic and stochastic, which makes PPO a perfect fit [87,88]. The problem is modeled as an MDP by a Device-to-Device (D2D)-assisted compute offloading approach based on PPO. Compared with three benchmarked techniques for 5G and beyond networks, it achieves superior latency and energy convergence by concurrently optimizing offloading decisions and resource allocation.

Task offloading has emerged as a critical optimization problem in vehicular edge computing (VEC) environments, and PPO has been increasingly used for this application [89]. In these, the vehicles, based on the computational overhead, dynamically decide on which of the following the tasks should be:

1. Executed locally;
2. Offloaded to nearby edge servers through Vehicle-to-Infrastructure (V2I) communication;
3. Delegated to neighboring vehicles through Vehicle-to-Vehicle (V2V) communication.

Vehicular networks are highly dynamic. Added to their mobility, there will be fluctuating wireless channel conditions and strict latency requirements. These situations make task-offloading decisions particularly challenging. Recent studies have demonstrated the effectiveness of PPO-based RL frameworks for addressing these challenges. Researchers proposed a PPO-based joint task-offloading and resource-allocation framework for vehicular edge computing [90]. The framework jointly optimizes communication and computational resources across V2I and V2V links, improving service latency and resource utilization. Similarly, PPO-based DRL approaches are used to deal with computation offloading in vehicular communications [54]. It demonstrated that adaptive policy learning can significantly improve task completion efficiency under dynamic traffic and network conditions. These studies further confirm PPO's suitability for communication–computation co-design problems involving continuous decision spaces, real-time adaptation requirements, and heterogeneous edge resources.

4.4. Integrated Sensing and Communication (ISAC)

In conventional communication networks, sensing (radar, localization, imaging) and communication (data transfer) are treated as separate processes that use different frequency bands. ISAC sees 6G waveforms as intrinsically containing both information bits and environmental-scattering information, thereby unifying both roles. [55]. ISAC is therefore a crucial component of 6G, as it reduces material resource consumption by eliminating the need for dedicated sensing hardware. In this case, each broadcast signal serves as both a radar probe and a data carrier. Due to growing spectrum scarcity, increased device connectivity, and its integration into an AI-ready network architecture, 6G prioritizes ISAC over 5G. [91,92]. AI-native 6G systems are expected to integrate sensing and computing capabilities directly into the network architecture, making ISAC a natural design paradigm. Such systems follow a layered approach [93], reduce the need for separate

sensing hardware and dedicated spectrum resources, and thereby lower overall material use [56]. Improved spectral efficiency may reduce infrastructure expansion requirements and associated deployment overhead in dense urban networks [8,94].

PPO supports the joint integration of sensing, computation, and control [95]. PPO-based approaches manage shared resources while preserving both sensing accuracy and communication reliability [96]. PPO is particularly suitable for ISAC because it can jointly optimize sensing and communication objectives within a unified policy framework. Its continuous-action capability enables adaptive adjustment of sensing and transmission parameters in dynamic environments. This flexibility helps maintain a balance between sensing accuracy and communication quality. This integration has also been applied to multi-agent formation control, where joint sensing-communication optimization reduces synchronization errors and closed-loop latency. Studies showed that such integration enables more reliable coordination in dynamic environments. Multiple automated guided vehicles performing cooperative tasks under base-station supervision [97] is an example of this. This model extends beyond conventional communication boundaries to represent integrated engineered systems.

4.5. UAV-Assisted Communications

UAVs and drones are part of the geospatial technology stack for surveillance, disaster response, and last-mile connectivity in urban and smart cities [98,99]. PPO-based energy management extends the operational endurance of UAVs. Researchers have extensively applied PPO to UAV communication systems, especially in areas that require real-time joint trajectory, power, and resource decisions [100]. PPO allows UAVs to learn long-term control strategies without requiring explicit mathematical models of the environment. The algorithm can simultaneously consider mobility, energy consumption, and communication quality in decision-making [101]. Such capabilities are valuable for autonomous UAV operations in rapidly changing network conditions. PPO-based DRL algorithm jointly controls 3D UAV positions, transmit power, and energy harvesting [102]. For secure UAV networks with eavesdropping, they can maximize a utility function that balances secrecy rate, energy efficiency, and interference. Such UAVs are commonly deployed in surveillance and defense applications. Similar enhanced policy optimization frameworks address physical layer security challenges in mobile communication systems under eavesdropping threats [51]. To capture both long-term mobility patterns and short-term channel fading, these algorithms use bidirectional long- and short-term memory methods. As a result, they maintain robustness against residual hardware defects and channel uncertainty while achieving improved secrecy-rate performance. Compared to fixed-deployment techniques, PPO achieves greater overall communication rates in STAR-RIS-assisted mobile wireless networks by concurrently optimizing hybrid beamforming and dynamically deploying STAR-RIS. A PPO-based sum rate maximization scheme for STAR-RIS-assisted vehicular networks underlaying UAV, achieving 22.05%, 35.12%, and 48.9% higher sum rate compared to DDPG, Zero forcing, and random schemes [103]. More coverage than traditional RIS is possible with Simultaneous Transmitting and Reflecting Reconfigurable Intelligent Surfaces (STAR-RIS), which can both broadcast and reflect signals.

Compared with off-policy RL techniques, the MAPPO strategy, which combines centralized training with decentralized execution, reduces the global Age-of-Updates (AoU) for UAV networks gathering time-sensitive IoT data by at least a factor of 1/2. MAPPO uses rate-splitting multiple access to maximize energy efficiency in multi-UAV wireless networks [58]. This approach jointly manages trajectory and beamforming, outperforming conventional multiple access schemes in complex environments. These results show that PPO effectively handles coupled mobility and communication constraints. Complement-

ing physical-layer optimizations, a logical network architecture requires flexible slicing mechanisms to effectively support diverse service requirements. In satellite-assisted UAV data-collection scenarios, a compound-action PPO variant has been used to jointly optimize UAV trajectory, sensor-node scheduling, and the choice between direct and satellite-relayed transmission, minimizing information staleness while accounting for energy and communication costs [104].

4.6. Network Slicing, Service Function Chaining, and Orchestration

Network slicing in 5G and B5G involves real-time or autonomous resource management to handle heterogeneous services [105]. Dynamic network slicing allows efficient sharing of physical infrastructure across multiple services [59]. It partitions a common network into multiple logical slices tailored to different service demands. This enables flexible resource allocation across core, radio access, and end-to-end network segments while supporting diverse vertical applications, heterogeneous user requirements, and tenant-based service models. As a result, existing infrastructure can be used without compromising network service quality. In network slicing, PPO dynamically partitions physical network resources into virtual slices tailored to specific service requirements such as eMBB and URLLC. PPO can be used for up/downlink time slots allocation in 5/6G Dynamic TDD networks, dynamically assigning resources based on changing user requests and QoS requirements [106]. The co-existence of eMBB and URLLC traffic within shared physical infrastructure creates a particularly acute scheduling conflict. This is because URLLC packets must preempt already-scheduled eMBB resources, and they do so within mini-slots. To address this, an RIS-assisted hybrid slicing framework has been proposed. This framework uses PPO to choose between two actions for each incoming URLLC packet: NOMA-based superposition or puncturing [107]. Compared to standalone NOMA and puncturing schemes, this hybrid approach achieves significant gains. It improves the URLLC packet admission rate by 5.31%. It also increases the eMBB sum rate by 144% over optimization-based baselines. Additionally, it reduces execution time by 78.3%. Complementing this, traffic-aware slice management has also been demonstrated in B5G O-RAN. This approach couples two components: a lightweight SVM-based KPI classifier and a PPO agent [108]. The PPO agent is used for slice-aware physical resource block allocation within the Near-RT RIC. Compared to static schedulers, this method achieves up to 25% higher eMBB throughput. It also achieves up to 28% lower URLLC latency. In addition, the system includes a silence-aware reclamation mechanism. This mechanism recovers more than 90% of idle PRBs when demand fluctuates. Enhanced Mobile Broadband (eMBB) supports high-data-rate services such as video streaming. In contrast, Ultra-Reliable Low-Latency Communication (URLLC) supports mission-critical applications that require very low latency and high reliability.

To achieve the broad goals, PPO-based frameworks have been developed for end-to-end network slicing in multi-layer MEC environments, especially for Computation Offloading and Resource Allocation [109,110]. Both independent and collaborative agents within these frameworks learn to maximize resource efficiency while satisfying per-slice QoS requirements [111]. Such deployments significantly outperform the standard greedy algorithms. Network slicing improves infrastructure utilization and resource efficiency [60].

For Service Function Chain (SFC) deployment, PPO effectively balances node deployment cost and end-to-end service latency while ensuring quality-of-service requirements [112,113]. To further enhance performance, complementary studies have integrated particle swarm optimization with DRL. They achieved faster convergence and lower deployment costs under strict resource constraints. Building on these advances, PPO was successfully applied to multi-objective network slice provisioning. In these, PPO approxi-

mates a solution to an online Multi-Objective Integer Programming problem. In the context of Industrial 5G, such AI-driven orchestration is becoming increasingly critical. It enables sustainable, adaptive network management [61]. This approach aligns with network slicing paradigms in which AI-driven coordination enables autonomous resource allocation, admission control, and scheduling across shared industrial infrastructure. Such intelligence is increasingly necessary in softwarized 5G, 6G, and beyond networks, where SDN, NFV, and edge computing support multiple industrial services with diverse, stringent quality-of-service requirements [61]. These solutions yield lower SLA violation rates and network operational costs than state-of-the-art solvers. Service Level Agreements (SLAs) define the minimum performance targets that network operators must meet, including latency, reliability, and throughput.

4.7. Satellite Communications

Efficient satellite resource allocation extends satellite operational life [62,114,115]. Studies evaluated PPO-based DRL architecture for joint power and bandwidth allocation in a 23-beam GEO multibeam satellite [116]. The work demonstrated superior performance, comparable to optimized genetic-algorithm approaches, while enabling substantially faster decision-making, suitable for real-time resource allocation. High-throughput satellites (HTS) with flexible multibeam architectures require dynamic resource allocation that is impractical with manual tuning [63]. Ultra-dense LEO deployment significantly introduces new challenges to connection continuity and network stability [64,117]. Addressing these challenges requires jointly optimizing multiple interdependent control variables rather than treating them in isolation. A dual-agent DRL framework was proposed that jointly optimizes satellite handover decisions and transmission power allocation in LEO networks [118]. By combining DQN for discrete handover control and PPO for continuous power allocation, the approach effectively handles mixed action spaces and improves coordination between coupled decision variables. Simulation results show notable gains in throughput and user service completion rates. They showcase the advantage of hybrid PPO-based designs over single-agent optimization strategies [119].

To address these problems, multi-agent RL frameworks for efficient handover management in integrated satellite-terrestrial networks are studied [120]. Under shared network resource constraints, a MAPPO-based handover strategy enables coordinated decision-making among several users. Combining centralized training with distributed execution reduces unnecessary handovers and failure rates. At the same time, they bring long-term benefits in dynamic LEO scenarios. Simulation results showed better QoS, with reduced drop effects and more stable mobility management, compared to conventional and single-agent learning approaches. Decoupling decision-making into temporal and spatial subproblems reduces the number of handovers and boosts success rates [64]. Such approaches show their effectiveness in highly dynamic LEO satellite environments with frequent handovers. The clipping mechanism in PPO prevents destabilizing policy updates that could cause catastrophic resource misallocation during demand peaks. Unmet System Demand (USD) denotes the portion of requested communication capacity that the satellite system cannot satisfy. Evaluations on a 23-beam GEO multibeam satellite show that a PPO-based policy achieves comparable USD performance to an optimized genetic algorithm while running roughly 1500 times faster, reaching a decision in about 0.08 s versus roughly 2 min for the genetic-algorithm baseline [116]—making DRL operationally feasible. Hierarchical RL frameworks dissociate handover timing and target selection for dense satellite constellations [64]. Temporal agents utilize PPO algorithms to determine handover timing by predicting future satellite load. Spatial agents effectively employ deep Q-networks to select target satellites based on utility functions. Apart from handover

and power management, PPO variants have also been applied to joint beam hopping and resource allocation in multi-LEO constellations, where hierarchical and multi-agent designs coordinate beam scheduling, bandwidth, and power across satellites to reduce load imbalance and inter-beam interference [121,122].

4.8. Vehicle-to-Everything (V2X) and Cognitive Radio

V2X communication optimization directly supports sustainable transportation, making it a key component of smart city planning [123,124]. V2X enables platooning and collision avoidance, enhances cooperative perception among connected vehicles, reduces their idle time, and lowers vehicular emissions [40]. It uses anti-jamming technology in dense urban traffic to improve traffic flow. It integrates well with roadside edge servers and blends with smart-city traffic management. V2X communications with energy harvesting can use a PPO-based multi-agent DRL framework [65]. Such systems use power splitting to divide harvested energy between information decoding and energy storage. Such implementations outperform traditional DRL resource allocation approaches in both effectiveness and robustness. This approach shows system-level innovation by jointly managing energy harvesting, communication decoding, and storage control.

PPO and its multi-agent variants are applied to cognitive radio and anti-jamming communication scenarios. In these, the agents learn adaptive spectrum access, power control, and interference mitigation strategies in dynamic adversarial environments. Recent studies have shown that MAPPO-based approaches are effective for cooperative jamming, cognitive-radio security, and intelligent anti-jamming decision-making in wireless communication systems [125–127]. Based on a single-agent PPO, MAPPO uses the KL-divergence and provides more stable data transmission than DQN in such complex vehicle environments [66]. The approach also supports computation offloading in vehicular edge networks, where it simultaneously manages task dependencies and stringent deadlines. Intelligent offloading decisions enabled by PPO reduce device battery drain and server energy usage. To achieve this, the algorithm uses surrogate clipping and generalized advantage estimation [12]. This helps achieve stable policy updates even in dynamic environments. This yields significant reductions in total delay and dropped-task ratios compared to DQN and related methods. These implementations demonstrate that PPO's stability and continuous control capabilities are effective across diverse network environments. Concurrently, complementary studies investigated opportunistic routing in urban environments to address mobility uncertainty [128]. By integrating convolutional neural networks with Q-learning, these approaches predict node transitions and optimize data-routing decisions across diverse urban scenarios [129].

4.9. Common Trends in PPO-Based Communication Systems

Across the reviewed studies, several common patterns emerge. Most PPO-based communication-system models formulate the state space using network-level observations. They include channel quality indicators, signal-to-interference-plus-noise ratio (SINR), queue lengths, traffic loads, user mobility information, resource-utilization levels, and energy-related metrics. Similarly, the action space is commonly defined through resource-management decisions. These decisions are mostly related to power-control adjustments, spectrum allocation, beam-selection strategies, routing decisions, computation offloading choices, and network slicing configurations.

Another notable trend is the design of reward functions that jointly optimize multiple performance objectives. The analysis and results focused on throughput maximization, latency reduction, energy efficiency, quality-of-service (QoS) enhancement, and resource-

utilization optimization. It reflects the inherently multi-objective nature of communication-network optimization problems.

Most studies rely on simulation-based evaluation using custom network simulators or established communication frameworks. This discusses the practical challenges of deploying and testing reinforcement-learning algorithms in operational communication infrastructures. The reported results consistently demonstrate improvements in network efficiency, reduced latency, optimized throughput, and improved energy management compared with conventional baseline methods. However, evidence of large-scale real-world deployment remains limited. Consequently, bridging the gap between simulation-based validation and practical deployment represents an important research direction for PPO-enabled communication systems.

4.10. Comparative Performance

Apart from raw performance measurements, sustainability metrics (particularly energy efficiency) are increasingly used as primary benchmarks in PPO-based communication studies. This shows a growing awareness of the environmental impact of intelligent network operation. As Table 5 shows, empirical studies consistently show that PPO offers an optimal balance between performance and stability, giving it an edge [13].

Table 5. Comparative Analysis of PPO-Based Studies across Communication System Applications.

Comparison	Communication Domain	Metric/Outcome	Representative Result
PPO vs. DQN	Vehicular edge computing	Total delay reduction	PPO reduced total delay by 13.85% compared with DQN under dynamic vehicle speeds and task deadlines [54].
PPO vs. DDQN	Vehicular edge computing	Total delay reduction	PPO reduced total delay by 11.24% compared with DDQN [54].
PPO vs. DDPG	RIS-assisted MU-MISO beamforming	Energy efficiency	PPO improved energy efficiency by 15.8% relative to DDPG [9].
PPO vs. Fractional Programming (FP)	RIS-assisted MU-MISO beamforming	Energy efficiency	PPO improved energy efficiency by 34.2% relative to FP [9].
PPO vs. Heuristic Algorithms	Heterogeneous networks (HetNets)	Throughput and user adaptation	PPO outperformed heuristic resource-allocation methods under varying user loads and channel conditions.
PPO vs. Greedy Algorithms	Network slicing and MEC	Resource efficiency/QoS	PPO-based slicing achieved better resource utilization and QoS satisfaction than greedy approaches [111].
PPO vs. Genetic Algorithm	GEO satellite power allocation	Runtime/unmet system demand	PPO achieved comparable unmet-demand performance while operating faster than an optimized genetic algorithm [116].
PPO vs. ACER	O-RAN resource allocation	Energy–latency trade-off	PPO achieved a more favorable balance between energy consumption and user latency than ACER.
MAPPO vs. IPPO	5G resource allocation/routing	Convergence speed, packet loss, and latency	MAPPO converged faster and achieved lower latency and packet loss than IPPO [30].

Note: Cross-study comparison should be interpreted cautiously because reviewed studies used different simulation environments, datasets, traffic models, network assumptions, and evaluation metrics.

Several reviewed studies reported measurable sustainability-related improvements associated with PPO-based optimization. Representative findings include a 15.8% increase in energy efficiency in RIS-assisted MU-MISO beamforming systems, a 34.2% improvement

over fractional programming baselines, and substantial reductions in delay and interference in vehicular and satellite communication environments. However, most studies evaluate operational efficiency rather than full lifecycle sustainability metrics. Metrics such as AI training energy consumption, inference cost, embodied carbon impact, and end-to-end deployment footprint remain underreported across the literature.

Despite these advantages, PPO typically requires greater computational resources and longer training times than lightweight heuristic methods. In some communication scenarios with limited training data or highly discrete action spaces, alternative DRL approaches may remain preferable.

Many reviewed studies reported that PPO outperforms value-based methods in continuous-control and high-dimensional optimization settings. This advantage was especially evident in problems involving continuous control variables, such as transmit power, beamforming, and computation offloading. Several studies observed improved performance relative to heuristic and optimization-based approaches under dynamic communication conditions. Energy-sensitive beamforming and satellite resource-allocation scenarios reported the largest gains. Compared with other DRL methods, PPO's main strengths were greater training stability and faster convergence in volatile network conditions.

However, PPO performance is highly dependent on the communication scenario, reward formulation, state representation, and hyperparameter calibration. Several reviewed studies used different simulation settings, traffic models, and evaluation metrics. This limits a direct cross-study comparison. PPO does not universally outperform alternative DRL methods in all conditions. For example, in highly discrete decision environments or sample-limited settings, value-based approaches such as DQN or DDQN may remain competitive. Similarly, PPO performance can degrade in certain situations. For example, in poorly designed reward structures, unstable channel observations, or insufficient exploration. In communication problems dominated by discrete decision spaces, such as channel selection or binary routing actions, value-based approaches like DQN and DDQN may offer lower computational complexity and simpler policy learning. Similarly, in sample-limited environments where large-scale policy training is impractical, off-policy methods may achieve faster learning efficiency. Lightweight heuristic or rule-based algorithms may also remain preferable in latency-sensitive edge deployments with limited computational resources. Therefore, the reported advantages are best interpreted within the context of the specific communication environment and evaluation framework used in each study.

4.11. *Synthesis of Research Questions*

The synthesis of literature provides clear and direct answers to the specific research questions formulated earlier in Section 3.1 of this review paper. Looking at the first question, PPO appears widely applied in resource allocation and beamforming across many different network types. Recurring design patterns for multi-agent coordination and hybrid architectures for energy-aware optimization in complex communication systems are documented in the literature. For the second question, future work faces challenges related to scalability and transitioning from simulation to real networks. Answers to the third question show PPO contributes directly to sustainable digital infrastructure through energy-efficient smart-city services and IoT. IoT coordination and intelligent communication management are key areas where these algorithms effectively support greener infrastructure goals in modern cities. Finally, this study identified several unresolved challenges (Section 5.1) and future directions (Section 6) for PPO-enabled communication systems that researchers should address soon.

5. Challenges and Limitations

5.1. Challenges with PPO Frameworks

PPO-based communication frameworks face substantial deployment challenges despite encouraging empirical findings. At an algorithmic level, it has limitations related to the static entropy coefficient and the hard-clipping mechanism, which require solutions [12]. The sim-to-real gap, the difficulty of reward structuring, multi-agent scalability, and the energy consumption and carbon footprint of AI-driven network operations are some of the major issues noted in the study. The majority of reviewed research does not assess PPO frameworks in practical smart-city or large-scale IoT scenarios; instead, they employ general wireless simulation setups. Therefore, there is a sim-to-real gap in the literature that has not been fully addressed, which limits the applicability of its conclusions to sustainable urban infrastructure environments.

Real-world channel dynamics, hardware impairments, and environmental interference make it difficult to replicate observations in synthetic testbeds faithfully. This gap undermines the practical transferability of simulation-trained models to operational wireless systems. Poorly calibrated models waste energy through unnecessary retransmissions and suboptimal power settings, making sim-to-real transfer a priority for both performance and efficiency.

The energy cost of AI training is another key challenge [130]. Training deep PPO models requires significant computational resources. Studies have noted that large-scale AI model training can require substantial GPU utilization and energy consumption, resulting in measurable carbon-equivalent emissions during training and deployment [130]. However, the carbon footprint and lifecycle energy cost of PPO-based reinforcement training remain underreported across communication-system studies [131]. Future evaluations should report standardized sustainability metrics, including GPU training hours, inference energy per decision cycle, communication overhead during distributed learning, carbon-equivalent emissions, and end-to-end operational energy savings during deployment. Reward shaping in multi-objective optimization is a challenge for AI models. It is still difficult to design reward functions that simultaneously balance throughput, latency, energy efficiency, and fairness. Poorly designed rewards may lead to unanticipated behavioral trade-offs or unsatisfactory convergence. Hybrid PPO-DQN systems see this challenge. To attain steady performance, they need careful multi-objective reward design. Similarly, although PPO is said to be simpler to tune, hyperparameter sensitivity persists. Careful calibration is still required for clipping parameters, learning rates, entropy coefficients, and GAE lambda values.

Misconfigured hyperparameters can destabilize training, particularly in fast-varying vehicular or satellite environments. Scalability and observability further complicate real-world adoption. While MAPPO demonstrates strong cooperative performance, scaling a large number of heterogeneous agents introduces significant coordination overhead. Most of the challenges will be related to credit assignment. MAPPO-based energy-saving control strategy can be used for ultra-dense network base stations. In such deployments, each base station will be treated as an agent requiring collaboration and interference management [132]. Integrated Adaptive Communication Network (IACN) built on MAPPO that dynamically adjusts communication topology and optimizes content for efficient multi-agent coordination [133]. Non-stationarity in ultra-dense 6G scenarios exacerbates these challenges.

5.2. Limitations of the Present Review

The review process identified substantial content overlapping across studies, particularly in RIS-assisted beamforming and satellite handover domains. Such redundancy

complicates synthesis and suggests a need for clearer differentiation of novel contributions. The study focused on the 2023–2026 timeframe to capture recent advances, but, in the process, might have necessarily excluded foundational pre-2023 works. This introduces some recency bias into the overall synthesis. Although the final set of 76 studies is sufficient for a focused review of recent PPO applications, some emerging work may not yet have appeared in peer-reviewed form or within the selected databases. Further, the analysis is limited to peer-reviewed studies and explicitly excludes preprints, industrial reports, and other unpublished results. This limitation matters because preprints often contain the latest work in fast-moving fields such as PPO, 6G, RIS, ISAC, and multi-agent reinforcement learning, several months before formal publication.

The review considered only English-language publications and omitted publications in Chinese, Korean, Japanese, and other languages. A large number of active research communities in the literature on wireless communications and AI were written in these languages [134,135]. Another limitation is the potential for publication bias. Studies reporting positive PPO performance improvements are more likely to be published than studies with negative, null, or inconclusive findings. As a result, the reviewed literature may overestimate the consistency of PPO's advantages across communication domains. Considerable heterogeneity in datasets, simulation settings, and performance metrics limits direct comparison across studies. Thus, determining whether reported performance gains arise from the PPO method itself or from differences in experimental setup is difficult to establish. Finally, most reviewed works rely on simulation-based validation, leaving real-world deployment evidence relatively limited.

A further limitation of the present review is its deliberate focus on PPO-based and related DRL approaches for communication-network optimization. Although PPO represents one of the most widely adopted DRL algorithms in recent 5G, B5G, and 6G research, other optimization paradigms include mathematical optimization, game-theoretic approaches, and heuristic and metaheuristic algorithms. Also, alternative DRL methods such as SAC, TD3, DDPG, TRPO, and A3C were not comprehensively analyzed. Therefore, the conclusions of this survey should be interpreted within the scope of PPO-centered research.

6. Future Research Directions

Future studies may extend the scope of this review. Researchers could conduct comprehensive comparative analyses of PPO and alternative optimization approaches, including SAC, TD3, DDPG, TRPO, A3C, heuristic methods, and mathematical optimization frameworks. Such investigations would provide deeper insights into the trade-offs among learning stability, sample efficiency, computational complexity, and scalability. They would also make practical deployment in next-generation communication networks easier. In addition, PPO applications are emerging in areas such as vehicular edge computing, distributed edge intelligence, and AI-native 6G networks. These remain promising directions for future research.

Real-World Validation and Sim-to-Real Transfer: Future research should quantify the energy cost of PPO training [136]. They need to consider inference pipelines and explore lightweight PPO variants [137] suitable for deployment on energy-harvesting IoT nodes and edge devices in smart cities. Literature indicates a growing need for advanced multi-agent reinforcement learning frameworks to address scalability and coordination challenges in next-generation communication systems. These studies should examine whether the significant computational overhead required to train the models could eventually undermine the net energy benefits achieved during their operational life.

Scalable Multi-agent PPO for Ultra-Dense Networks: Recent dual-clip MAPPO implementations in energy-constrained aerial systems demonstrate improved convergence

and coordination stability [138]. Convergence speeds during trajectory optimization improve significantly when employing adaptive discounting strategies. Future work should extend these design principles to heterogeneous multi-tier network architectures.

Semantic-aware communication reduces unnecessary data transmission, lowering energy consumption and enabling sustainable bandwidth utilization in IoT-heavy smart city networks. Integrating PPO-based resource control with adaptive, knowledge-driven mechanisms can enable efficient handling of semantic redundancy and improve system-level quality of experience. These developments point toward unified cross-layer optimization strategies that bridge semantic processing and physical-layer design in next-generation communication systems. Building upon foundational machine learning successes in maritime networks, future frameworks should incorporate deeper semantic awareness across communication layers [46]. Future investigations must prioritize hybrid architectures that balance computational efficiency with real-time decision-making capabilities in offshore environments.

7. Implications for Sustainable and Smart Infrastructure

7.1. Sustainability and Practical Implications

The review has direct implications for modernizing digital communication systems and infrastructure. Integrating sensing and communication functions can reduce dedicated hardware requirements and associated spectrum overhead in large IoT deployments [56,93]. Future communication architectures should jointly optimize throughput, latency, and operational energy consumption while accounting for the computational overhead of AI-driven control. Theoretical models should incorporate carbon footprint alongside throughput and latency as primary design objectives.

Network designers who are planning to create new communication system architectures in the near future can prioritize energy efficiency. Future 5G and 6G deployment strategies must explicitly account for algorithmic energy costs [92]. It is expected that collaboration between industry and academia will steadily improve the deployment of AI-driven solutions.

7.2. Implications for Smart Infrastructure

PPO effectively manages the spectrum during peak loads, making urban communication systems more resilient [139]. Adaptive public infrastructure will benefit from self-optimizing networks that automatically adjust to changing citizen demands. Such infrastructure will be a key component of adaptive governance, especially in smart cities, which, in turn, help ensure proper service delivery [140]. Sustainable transportation systems use V2X technologies to reduce congestion, significantly reduce delays and energy use, and thereby lead to positive environmental and economic impacts [141].

8. Conclusions

Across 76 peer-reviewed studies published between 2023 and 2026, PPO has emerged as one of the most effective deep reinforcement learning approaches for optimizing next-generation communication systems. Significant improvements are consistently reported in energy efficiency, spectral utilization, and system-level coordination across resource allocation, beamforming, routing, MEC, ISAC, UAV-assisted networks, satellite communications, and V2X systems. Across these domains, PPO performs particularly well in environments characterized by continuous control, dynamic network conditions, and competing optimization objectives.

Several limitations remain. The sim-to-real transfer gap continues to hinder practical deployment because operational network conditions are difficult to reproduce accurately

in simulation environments. Reward design for multi-objective optimization remains challenging, while MAPPO-based coordination becomes increasingly complex as agent density scales toward ultra-dense 6G infrastructures.

Future research must prioritize real-world testbeds, standardized sustainability metrics, and reproducible multi-agent evaluation frameworks. The convergence of 6G, edge intelligence, and large-scale IoT environments creates conditions where PPO-based optimization is especially valuable, given the high-dimensional state spaces, heterogeneous agents, continuous action requirements, and strict energy constraints. PPO is steadily evolving from a task-specific optimization technique into a foundational control framework for intelligent, adaptive, and energy-aware communication ecosystems.

Author Contributions: Conceptualization, V.K.M. and T.T.; methodology, B.M.; formal analysis, V.K.M.; validation, B.M. and T.T.; investigation, V.K.M.; resources, T.T.; data curation, T.T.; writing—original draft preparation, V.K.M.; writing—review and editing, T.T.; visualization, B.M.; supervision, B.M.; project administration, V.K.M.; All authors contributed to manuscript conception, literature synthesis, and to the sustainability framing of PPO applications. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing does not apply to this article.

Acknowledgments: We acknowledge the academic support given by Nagajyothi Koripella of BEST Innovation University, India, during the development of this research. Tools such as Grammarly (v1.8.0) and Scopus AI are used in research, primarily to correct spelling mistakes (typos), fix grammar, and revalidate claims made in the work. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

5G	Fifth Generation
6G	Sixth Generation
B5G	Beyond Fifth Generation
A3C	Advantage Actor–Critic
ABR	Adaptive Bitrate
ACER	Actor–Critic with Experience Replay
AI	Artificial Intelligence
AoI	Age of Information
AoU	Age of Updates
CD-PPO	Contribution-based Dual-clip PPO
CTDE	Centralized Training with Decentralized Execution
CRN	Cognitive Radio Network
D2D	Device-to-Device
DDPG	Deep Deterministic Policy Gradient
DL	Deep Learning
DPPO	Distributed Proximal Policy Optimization
DRL	Deep Reinforcement Learning
DSA	Dynamic Spectrum Access
DNN	Deep Neural Network

ECN	Explicit Congestion Notification
eMBB	Enhanced Mobile Broadband
FL	Federated Learning
FWA	Fixed Wireless Access
GAN	Generative Adversarial Network
GAT	Graph Attention Networks
GAE	Generalized Advantage Estimation
GEO	Geostationary Earth Orbit
HTS	High-Throughput Satellite
IoT	Internet of Things
IoV	Internet of Vehicles
IPPO	Independent Proximal Policy Optimization
IRS	Intelligent Reflecting Surface
ISAC	Integrated Sensing and Communication
ISCC	Integrated Sensing, Communication, and Computing
KL	Kullback–Leibler
LEO	Low Earth Orbit
LSTM	Long Short-Term Memory
MAPPO	Multi-agent Proximal Policy Optimization
MARL	Multi-agent Reinforcement Learning
MEC	Multi-Access Edge Computing
MDP	Markov Decision Process
MIMO	Multiple-Input Multiple-Output
MISO	Multiple-Input Single-Output
ML	Machine Learning
mmWave	Millimeter Wave
MOS	Mean Opinion Scores
O-RAN	Open Radio Access Network
PPO	Proximal Policy Optimization
QoE	Quality of Experience
QoS	Quality of Service
RIS	Reconfigurable Intelligent Surface
RL	Reinforcement Learning
RSMA	Rate-Splitting Multiple Access
SAC	Soft Actor-Critic
SCA	Successive Convex Approximation
SFC	Service Function Chaining
SLA	Service Level Agreement
SNN	Spiking Neural Networks
STAR-RIS	Simultaneous Transmitting and Reflecting Reconfigurable Intelligent Surface
SU	Secondary User
SWIPT	Simultaneous Wireless Information and Power Transfer
TD3	Twin Delayed DDPG
THz	Terahertz
TRPO	Trust Region Policy Optimization
UAV	Unmanned Aerial Vehicle
URLLC	Ultra-Reliable Low-Latency Communication
USD	Unmet System Demand
V2X	Vehicle-to-Everything
WP-IoT	Wireless Powered Internet of Things

Appendix A

Appendix A presents the studies included in the qualitative synthesis of this systematic review. The table summarizes peer-reviewed studies in which PPO, MAPPO, IPPO, DPPO, or another named PPO variant was the primary or comparator algorithm, and provides the primary communication domain, PPO variant, and main research contribution for each. Studies are included in this table only where the cited source itself implements or analyzes a PPO-family algorithm. The PPO Variant column identifies the specific PPO variant employed in each study whenever explicitly reported by the authors. In many cases, studies adopted the standard single-agent PPO algorithm without introducing or naming a specialized variant. Therefore, in such cases, the studies are classified simply as “PPO”. Studies employing extensions such as MAPPO, IPPO, DPPO, PPO-Transformer, PPO-DQN, and other PPO-derived architectures are identified using their corresponding variant names. This approach ensures that studies are classified according to the algorithm actually implemented or analyzed rather than broader framework descriptions, application domains, or architectural characteristics. General reviews, surveys, and broader DRL/MARL studies that discuss PPO only in passing are cited in the main text instead. This appendix improves the transparency, traceability, and reproducibility of the review methodology. Other references cited throughout the manuscript, including surveys, reviews, and contextual literature, were used for background purposes and were not part of the reviewed study set.

Table A1. PPO and Related Communication-System Studies Included in the Review.

Ref. No.	Authors	Year	Communication Domain	PPO Variant/Study Type	Contribution
[9]	Eskandari et al.	2024	RIS-aided MU-MISO Systems	PPO	Proposes a joint beamforming algorithm using statistical CSI and the PPO algorithm to maximize the ergodic sum rate while significantly reducing channel estimation overhead compared to instantaneous CSI methods.
[14]	Sheikh et al.	2026	Smart City Systems/System-of-Systems Networks	MAPPO/PPO-based MARL Framework	Multi-agent reinforcement learning framework for optimizing smart-city communication coordination and infrastructure management
[27]	Shen et al.	2025	Tactical Communication Networks	PPO + GRU + Graph Attention	Intelligent path selection
[28]	Zhang et al.	2023	RIS-Assisted SWIPT Networks with RSMA	PPO	PPO-based DRL framework to jointly optimize transmit beamforming vectors, power splitting (PS) ratios, common message rates, and RIS phase shifts in unison to maximize energy efficiency.
[30]	Viana et al.	2025	UAV Communications	MAPPO-Transformer	Secure UAV communication and resilience optimization. Integrates MAPPO with Transformer-based detection for resilient UAV links.
[31]	Zuo et al.	2025	6G Space-Air-Ground Integrated Network (SAGIN)	MAPPO	Shows that a multi-agent PPO architecture can effectively coordinate communication-network decisions under realistic network constraints
[35]	Hong et al.	2026	Wireless Channel Access	IPPO (Specifically quantized QIPPO/CA)	Distributed channel-access optimization with quantized communication. Introduces a quantized IPPO framework (QIPPO/CA) for efficient channel access.

Table A1. Cont.

Ref. No.	Authors	Year	Communication Domain	PPO Variant/Study Type	Contribution
[36]	Wang et al.	2025	Cognitive Radio Networks	MAPPO	Dynamic spectrum access in heterogeneous wireless systems. Employs the HAPPO algorithm for heterogeneous-agent spectrum access. Automatic ECN tuning scheme based on the IPPO algorithm that uses a decentralized training and execution paradigm to handle in-cast congestion and mixed mice–elephant traffic patterns.
[37]	Wang et al.	2025	High-Speed Data Center Networks (DCNs)	IPPO	Constrained multi-agent DRL approach using JointGAT (Graph Attention Network) arc
[38]	Lin et al.	2026	Probabilistic Routing/Traffic Engineering	MAPPO (Specifically JGAT-MAPPO)	CD-PPO algorithm to jointly optimize subchannel selection and power allocation, maximizing intra-platoon transmission success ratios and V2I Mean Opinion Scores (MOS).
[40]	Xu et al.	2023	Vehicular Networks	Contribution-based Dual-clip PPO	Develops a multi-tier offloading framework using PPO to intelligently allocate vehicular tasks across edge, regional, and cloud layers based on real-time feedback like congestion and CPU utilization.
[41]	Alsahfi et al.	2025	Vehicular Big Data (VBD) Offloading	PPO	Adaptive resource allocation in A2G-TN systems. Introduces CL-MAPPO for cross-layer resource block allocation. Integrates energy-efficient Spiking Neural Networks (SNNs) into a PPO-based Actor-Critic framework to jointly optimize beamforming and power allocation for integrated sensing and communication.
[42]	Rehman et al.	2025	Cross-layer 6G Systems/6G A2G-TN Systems	MAPPO	An optimization framework for energy-harvesting vehicular communication networks, and explicitly evaluates the Age of Information (AoI) performance
[43]	Shang et al.	2026	ISAC-assisted V2X Networks	SNN-driven PPO (Spiking Actor-Critic PPO)	Adaptive transmission in nonstationary environments. Develops a PPO-based adaptive mode and modulation selection scheme for dynamic channels.
[45]	Wang et al.	2024	V2X/Integrated Sensing/Resource Optimization	PPO	Analyzes MAML-PPO as a comparator in a knowledge-embedded resource allocation study
[47]	Lin et al.	2025	MIMO Systems	PPO	Implements a PPO-based intelligent resource allocation mechanism to maximize device energy efficiency and satisfy dynamic traffic demands within vehicle cabins.
[49]	Mao et al.	2025	Maritime Wireless Networks	PPO/MetaRL	Secure communication via enhanced PPO. Proposes PPO-BOP incorporating BiLSTM and off-policy feedback for secrecy rates
[50]	Lin et al.	2023	In-Vehicle Heterogeneous Networks (HetNets)	PPO	Collaborative inference optimization. Introduces the MAHPPO multi-agent framework to optimize DNN partitioning and resource scheduling.
[51]	Ma et al.	2026	Secure Mobile Communications	PPO-BiLSTM	
[52]	Fang et al.	2026	IoT Edge Systems/Computing	MAPPO	

Table A1. Cont.

Ref. No.	Authors	Year	Communication Domain	PPO Variant/Study Type	Contribution
[54]	Mustafa et al.	2025	Vehicular Edge Computing	PPO	PPO-based algorithm using Generalized Advantage Estimation (GAE) and surrogate clipping to optimize offloading decisions, minimizing delays and task drop ratios in dynamic vehicular networks
[58]	Wang et al.	2026	UAV Communication/ Multi-UAV Wireless Networks	MAPPO/DRL-based RSMA Framework	Energy-efficient RSMA optimization in UAV-assisted wireless-powered communication networks. EEMACO algorithm based on MAPPO to maximize weighted sum rates.
[64]	Zhao et al.	2025	Satellite Communications/LEO Satellite Networks	Hybrid PPO-DQN	Handover and power allocation optimization/Proposes a hierarchical framework using PPO for timing and DQN for location in handovers.
[65]	Marzuk et al.	2025	V2X/O-RAN	DRL	Proposes an optimized distributed computation offloading (ODCO) framework using PPO to minimize latency and energy consumption in 5G/6G O-RAN-based V2X networks.
[67]	Su et al.	2026	Dense 5G Planning	Hierarchical MAPPO	Scalable network optimization/Proposes HMAPPO-RL to optimize base station placement and antenna beamwidth
[68]	Sun et al.	2025	UAV Networks/Resource Allocation	Hierarchical PPO (empirical simulation)	Hierarchical PPO for dynamic resource allocation in UAV networks, demonstrating improved spectrum efficiency and service quality.
[70]	He et al.	2025	O-RAN/Digital Twin	HAPPO	Heterogeneous-Agent PPO RL for xApps Coordination in Digital Twin Enabled O-RAN
[71]	Qazzaz et al.	2026	6G O-RAN/Green Communications	PPO	A hierarchical rApp-xApp framework that jointly optimizes radio unit activation and multi-criteria user association weights to reduce power
[72]	Asemian et al.	2026	MEC-O-RAN/Anti-Jamming	PPO (Jamming Estimator)	DHRL framework using a PPO-based jamming estimator (ADPHRP-JE) to predict jammed slots and a Transformer-based task scheduler
[74]	Lee and Kim	2024	Hybrid RL Optimization	Hybrid PPO-DQN	Improved exploration and policy stability. Implements a dual-agent framework (PPO-DQN variant) for safety-critical navigation.
[78]	Luo et al.	2024	Video Streaming (ABR)	BC-PPO (Behavior Cloning + PPO)	Proposes BC-PPO ABR to address slow convergence in learning-based ABR algorithms, using PPO to handle severe network fluctuations in mmWave 5G environments.
[79]	Iqbal et al.	2026	RIS-Assisted Multi-User MISO Systems	PPO	Develops an on-policy PPO-based DRL algorithm to jointly optimize base station beamforming and RIS phase shifts, significantly reducing computational complexity compared to Fractional Programming (FP).
[80]	Wara et al.	2025	Full-Duplex RIS-Aided NOMA-ISAC	MAPPO	Uses Centralized Training with Decentralized Execution (CTDE) to maximize minimum beam pattern gain by jointly controlling beamforming, RIS configuration, and power allocation.

Table A1. Cont.

Ref. No.	Authors	Year	Communication Domain	PPO Variant/Study Type	Contribution
[81]	Hu et al.	2024	Secure MmWave D2D Networks	Nested PPO/MAPPO	Employs a nested DRL structure using discrete PPO for RIS-user association and MAPPO for multi-agent
[87]	Xie et al.	2025	UAV-Assisted Semantic D2D Networks	GNN-enabled PPO	Integrates heterogeneous Graph Neural Networks with PPO to jointly optimize MU transmission power, channel allocation, and semantic symbol rates to maximize QoE under malicious jamming. Dynamic partitioning scheme for idle/active devices using MAPPO to minimize long-term average task delay for delay-sensitive applications under strict deadline constraints.
[88]	He et al.	2024	D2D Mobile-Edge Computing (MEC)	MAPPO	Jointly optimizes beamforming and phase shifts using PPO with surrogate clipping to maximize bit-per-joule energy efficiency in MU-MISO systems.
[82]	Iqbal et al.	2024	RIS-Assisted MU-MISO Systems	PPO/DPPO	MAPPO for joint caching, computation, and resource management in UAV-vehicular MEC.
[83]	Wu et al.	2025	UAV-Assisted Vehicular MEC	MAPPO (joint caching/computation)	Enhanced MAPPO for cooperative dynamic target search in multi-UAV systems.
[86]	Zhang et al.	2025	Multi-UAV Cooperative Systems	AS-MAPPO (improved multi-agent PPO)	Utilizes an improved MAPPO algorithm to decouple centralized training from distributed execution and a server-weighted scoring selection (SS) algorithm to optimize task offloading across cloud-edge-device collaborative layers while balancing Quality of Experience (QoE) and energy consumption.
[89]	Hu et al.	2026	6G-enabled IoV/VEC	MAPPO (Improved for CTDE decoupling)	Proposes a computationally robust PPO algorithm to optimize joint communication, perception, and task offloading in edge-assisted ISAC systems under computation uncertainty.
[95]	Shen et al.	2024	Edge Computing/ISAC	PPO (with Robust Design)	MAPPO-driven resource allocation for ISAC in UAV/LEO scenarios.
[96]	Sun et al.	2025	ISAC	MAPPO	Solves the multi-agent networked formation control problem using MAPPO, leveraging global ISCC state information to reduce synchronization errors and latency
[97]	Zhou et al.	2026	Industrial Wireless/Formation Control	Multi-Agent PPO	PPO-based DRL agent with a multi-action space to simultaneously optimize UAV 3D trajectory, transmit power, IoT node association, and power allocation factors to balance energy efficiency and far-near fairness
[100]	Ghomri et al.	2024	NOMA-UAV Networks/IoT	PPO	Utilizes PPO for its sample efficiency and stability to minimize total energy consumption by jointly optimizing task offloading, aerial STAR-RIS trajectory, amplitude and phase shift coefficients, and power allocation
[101]	Aung et al.	2024	Aerial STAR-RIS-assisted MEC	PPO	

Table A1. Cont.

Ref. No.	Authors	Year	Communication Domain	PPO Variant/Study Type	Contribution
[102]	Wang et al.	2024	Multi-UAV Relay Communication	Mix-Greedy MAPPO	MAPPO for path planning and relay communication in air-ground UAV networks.
[103]	Chaudhary et al.	2025	STAR-RIS-assisted V2V Networks	PPO	Implements a PPO-based DRL algorithm to maximize system sum rates in STAR-RIS networks by jointly optimizing beamforming vectors, coefficient matrices, and symbol rates.
[104]	Zhou et al.	2024	IoRT/UAV-Satellite Integrated Networks	Compound-action PPO (CPPO)	CPPO to handle mixed continuous and discrete action spaces, optimizing UAV trajectories, sensor scheduling, and transmission decisions to balance AoI, energy, and costs
[106]	Nguyen & Kim	2025	5G/6G Dynamic TDD Networks	PPO-TA (Actor-Critic PPO variant)	PPO-TA dynamically schedules TDD time slots to maximize the uplink/downlink sum rate using a KL-divergence-penalized objective function to ensure training stability and QoS compliance.
[107]	Adhikari et al.	2025	6G/RIS-Assisted Hybrid Slicing	PPO	Implements a hybrid slicing technique (NOMA + puncturing) with PPO to intelligently schedule URLLC traffic on top of eMBB traffic
[108]	Raja et al.	2026	B5G O-RAN/Traffic Classification	PPO	Couples SVM-based classification with a PPO agent for slice-aware PRB allocation, including a silence-aware resource reclamation mechanism
[109]	Zhang et al.	2024	MEC/D2D Communication	PPO (Actor-Critic PPO)	Jointly optimizes task offloading and resource allocation in D2D-assisted MEC, outperforming DQN and A2C in balancing delay and energy consumption.
[110]	Hikmat & Sahib et al.	2026	Network Slicing, Service Function Chaining, and Orchestration	PPO-MDP	Dynamic resource allocation and network slicing in 5G, with empirical gains in throughput, energy efficiency, fairness, and QoS compared to traditional methods (GA, PSO, etc.).
[114]	Fu et al.	2025	RIS-assisted NOMA Satellite Networks	PPO-DQN	Merges PPO (for continuous phase-shift and power variables) and DQN (for discrete user pairing) to maximize the average sum rate in 6G satellite-ground communications
[115]	Zhang et al.	2026	LoRa-based Direct-to-Satellite IoT Networks	Self-Attention PPO (SAPPO)	Integrates a self-attention mechanism into the PPO algorithm to optimize uplink scheduling and fairness while avoiding packet collisions in LEO satellite IoT environments
[116]	Xu et al.	2023	Multibeam GEO Satellite Communications	PPO	Joint power and bandwidth allocation using PPO-based DRL; achieved USC performance comparable to optimized genetic algorithms with substantially lower computation time.
[117]	Li et al.	2025	Ultra-Dense LEO Satellite Networks/Packet Routing	MAPPO	MAPPO-based routing algorithm integrated with Graph Attention Networks (GATs) and an M/M/1/K queuing model to minimize communication delay and energy consumption across dynamic topologies.

Table A1. Cont.

Ref. No.	Authors	Year	Communication Domain	PPO Variant/Study Type	Contribution
[118]	Zhang et al.	2026	Satellite Communications/LEO Satellite Networks	Hybrid DQN-PPO/Dual-Agent PPO	Joint optimization of satellite handover and power allocation in LEO satellite systems
[119]	Sadiki et al.	2023	MIMO-based Multi-access Edge Computing (MEC)	PPO	Formulates the offloading problem in a massive MIMO-MEC system as an MDP and introduces a PPO-based algorithm to solve the limitations of discrete action spaces, specifically for continuous power allocation.
[120]	Li et al.	2026	Satellite-Terrestrial Integrated Networks (STINs)	MAPPO	Proposes a handover-oriented learning scheme for multi-user STINs using MAPPO under a centralized training Digital Twin-empowered framework
[121]	Zhao et al.	2025	Multi-LEO Satellite Networks	Hierarchical PPO (HPPO)	using Hierarchical PPO (HPPO) to jointly optimize beam hopping, bandwidth, and power allocation, effectively reducing satellite load imbalances.
[122]	Meng et al.	2025	LEO Satellite Communication/Hybrid Wide-Spot Beam Coverage	MAPPO	Cooperative MAPPO algorithm to jointly optimize power allocation and dynamic beam hopping to maximize throughput and minimize delay fairness among beam positions
[123]	Alharbi	2025	Smart City Systems	Deep Multi-Objective PPO	Review of deep multi-objective reinforcement learning and vision-based systems for smart cities. PPO as a key algorithm for balancing conflicting urban goals like congestion and energy management
[124]	Louati et al.	2024	Autonomous Vehicle Networks	Multi-Agent PPO	Cooperative autonomous vehicle coordination for sustainable smart city environments. Benchmark a novel Multi-Agent Actor-Critic (MA2C) algorithm against MAPPO for multi-AV lane-changing decisions, emphasizing passenger comfort and energy efficiency gains across varying traffic densities.
[127]	Zhang et al.	2025	Wireless Security	MAPPO	Proposes a multi-user anti-jamming algorithm using MAPPO for throughput optimization
[132]	Zhen et al.	2025	5G Ultra-Dense Networks (UDN)	MAPPO + MSD	The base station control algorithm uses MAPPO, integrated with a novel Mode Switching Decision (MSD) algorithm that uses cosine similarity to minimize unnecessary sleep-mode transitions and energy waste.
[133]	Wang et al.	2025	Multi-Agent Navigation Systems	MAPPO	Proposes the Integrated Adaptive Communication Network (IACN) based on MAPPO, featuring dynamic topology adjustment via learnable graphs, content optimization for task relevance, and adaptive frequency modulation via Bayesian Networks.

Table A1. Cont.

Ref. No.	Authors	Year	Communication Domain	PPO Variant/Study Type	Contribution
[136]	Espinosa et al.	2025	Edge-Cloud ECC/V2X	PPO	Performs a comparative study between PPO and Particle Swarm Optimization (PSO) for energy-efficient task allocation in Kubernetes-orchestrated clusters, identifying PPO as a faster and more computationally lightweight option for resource-constrained devices.
[137]	Wu et al.	2025	Vehicular Edge Computing	PPO-Transformer (AHP-PPO)	Lightweight adaptive task offloading optimization in vehicular edge computing. PPO with a lightweight Transformer using dual-stream attention to optimize joint task offloading and proactive caching in dynamic IoV environments.
[138]	Li et al.	2025	Air-Ground Vehicular Edge Computing Networks	MAPPO/DRL-based PPO Framework	Joint task offloading and resource allocation in energy-harvesting vehicular edge networks. DC-MAPPO (Dual-Clip MAPPO) for task offloading in air-ground networks.
[139]	Wu et al.	2026	IoT Networks/Dynamic Spectrum Access (DSA)	Decentralized PPO	Fully decentralized framework that uses an interference-aware state-action representation and adaptive reward shaping to optimize throughput and reduce MAC-layer latency in dense IoT environments
[142]	Liu et al.	2026	mmWave Systems	PPO-SelfAttention	Proposes SAPPO to enhance beam tracking accuracy via self-attention mechanisms.
[143]	Liu et al.	2025	UAV-assisted ISAC	PPO	Implements PPO within a Federated Learning framework for ISCC resource optimization.
[144]	Hu et al.	2025	Wireless Powered IoT (WP-IoT)/AoI	HVF-based PPO (Hybrid Value Function PPO)	Integrates clipped and unclipped value functions to jointly optimize scheduling and power control for minimizing the average Age of Information (AoI) in hybrid-action spaces.
[145]	Chen et al.	2026	Semantic Communications	Hybrid PPO-DDPG	Integrates PPO for power management and DDPG for knowledge base adaptation in RSMA systems.

References

- Ericsson. EMR June 2025 Highlights Growing Monetization Appeal of 5G Fixed Wireless Access. 2025. Available online: <https://www.ericsson.com/en/press-releases/2025/6/emr-june-2025-highlights-growing-monetization-appeal-of-5g-fixed-wireless-access> (accessed on 24 June 2026).
- Statista. 5G—Statistics & Facts. 2025. Available online: <https://www.statista.com/topics/3447/5g/?srsltid=AfmBOoqwwLz1PwaNuxcOGXpe4chSvyttaEnxiMFzNriHlbev7lnechxh-> (accessed on 24 June 2026).
- Hoang, D.T.; Huynh, N.V.; Nguyen, D.N.; Hossain, E.; Niyato, D. *Deep Reinforcement Learning for Wireless Communications and Networking: Theory, Applications and Implementation*, 1st ed.; Wiley: Hoboken, NJ, USA, 2023.
- Chen, A.C.H.; Jia, W.-K.; Hwang, F.-J.; Liu, G.; Song, F.; Pu, L. Machine Learning and Deep Learning Methods for Wireless Network Applications. *EURASIP J. Wirel. Commun. Netw.* **2022**, *2022*, 115. [[CrossRef](#)]
- Wang, Y.; Lei, J.; Shang, F.; Li, Y. A Comprehensive Survey of Multi-Agent Deep Reinforcement Learning for Wireless Spectrum Management. *Neurocomputing* **2025**, *653*, 131236. [[CrossRef](#)]
- Goel, A.; Masurkar, S.; Pathade, G.R. An Overview of Digital Transformation and Environmental Sustainability: Threats, Opportunities, and Solutions. *Sustainability* **2024**, *16*, 11079. [[CrossRef](#)]

7. Safitra, M.F.; Lubis, M.; Kurniawan, M.T.; Alhari, M.I.; Nuraliza, H.; Azzahra, S.F.; Putri, D.P. Green Networking: Challenges, Opportunities, and Future Trends for Sustainable Development. In *Proceedings of the 2023 11th International Conference on Computer and Communications Management, Nagoya Japan, 4–6 August 2023*; ACM: New York, NY, USA, 2023; pp. 168–173.
8. Kumar, R.; Gupta, S.K.; Wang, H.-C.; Kumari, C.S.; Korlam, S.S.V.P. From Efficiency to Sustainability: Exploring the Potential of 6G for a Greener Future. *Sustainability* **2023**, *15*, 16387. [\[CrossRef\]](#)
9. Eskandari, M.; Zhu, H.; Shojaeifard, A.; Wang, J. Statistical CSI-Based Beamforming for RIS-Aided Multiuser MISO Systems via Deep Reinforcement Learning. *IEEE Wirel. Commun. Lett.* **2024**, *13*, 570–574. [\[CrossRef\]](#)
10. Zhu, C.; Dastani, M.; Wang, S. A Survey of Multi-Agent Deep Reinforcement Learning with Communication. *Auton. Agents Multi-Agent Syst.* **2024**, *38*, 4. [\[CrossRef\]](#)
11. Zheng, M.; Zhang, J.; Zhan, C.; Ren, X.; Lü, S. Proximal Policy Optimization with Reward-Based Prioritization. *Expert Syst. Appl.* **2025**, *283*, 127659. [\[CrossRef\]](#)
12. Sha, S.; Liu, Y.; Huo, B. Dynamic Proximal Policy Optimization: Enhancing PPO with Adaptive Entropy and Smooth Clipping. *Neurocomputing* **2026**, *674*, 132861. [\[CrossRef\]](#)
13. Cui, H. Evaluating the Performance Metrics of PPO, DQN, and DDPG in Continuous Control Tasks. *ITM Web Conf.* **2025**, *78*, 01009. [\[CrossRef\]](#)
14. Sheikh, A.; Chong, E.K.P. Multi-Agent Reinforcement Learning Framework for Optimizing Smart Cities as System of Systems. *Syst. Eng.* **2026**, *29*, 3–19. [\[CrossRef\]](#)
15. Cheng, P.; Chen, Y.; Ding, M.; Chen, Z.; Liu, S.; Chen, Y.-P.P. Deep Reinforcement Learning for Online Resource Allocation in IoT Networks: Technology, Development, and Future Challenges. *IEEE Commun. Mag.* **2023**, *61*, 111–117. [\[CrossRef\]](#)
16. Jiao, L.; Shao, Y.; Sun, L.; Liu, F.; Yang, S.; Ma, W.; Li, L.; Liu, X.; Hou, B.; Zhang, X.; et al. Advanced Deep Learning Models for 6G: Overview, Opportunities, and Challenges. *IEEE Access* **2024**, *12*, 133245–133314. [\[CrossRef\]](#)
17. Ericsson. Mobile Subscriptions Outlook. Available online: <https://www.ericsson.com/en/reports-and-papers/mobility-report/dataforecasts/mobile-subscriptions-outlook> (accessed on 24 June 2026).
18. Markets and Markets 6G Market Size & Outlook, 2030–2036. 2025. Available online: <https://www.marketsandmarkets.com/Market-Reports/6g-market-213693378.html> (accessed on 24 June 2026).
19. Haritwal, S.; Baul, S. *6G Market*; NMSC: Boston, MA, USA, 2025.
20. Latreche, S.; Bellahsene, H. A Comprehensive Survey on 6G: Enabling Technologies, Key Applications, and Future Challenges. *Frankl. Open* **2026**, *15*, 100559. [\[CrossRef\]](#)
21. Fernando, X.; Lăzăroi, G. Energy-Efficient Industrial Internet of Things in Green 6G Networks. *Appl. Sci.* **2024**, *14*, 8558. [\[CrossRef\]](#)
22. Musaddiq, A.; Olsson, T.; Ahlgren, F. Reinforcement-Learning-Based Routing and Resource Management for Internet of Things Environments: Theoretical Perspective and Challenges. *Sensors* **2023**, *23*, 8263. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Puspitasari, A.A.; Lee, B.M. A Survey on Reinforcement Learning for Reconfigurable Intelligent Surfaces in Wireless Communications. *Sensors* **2023**, *23*, 2554. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Ismail, A.A.; Khalifa, N.E.; El-Khoribi, R.A. A Survey on Resource Scheduling Approaches in Multi-Access Edge Computing Environment: A Deep Reinforcement Learning Study. *Clust. Comput.* **2025**, *28*, 184. [\[CrossRef\]](#)
25. Hady, M.A.; Hu, S.; Pratama, M.; Cao, Z.; Kowalczyk, R. Multi-Agent Reinforcement Learning for Resources Allocation Optimization: A Survey. *Artif. Intell. Rev.* **2025**, *58*, 354. [\[CrossRef\]](#)
26. Amodu, O.A.; Althumali, H.; Mohd Hanapi, Z.; Jarray, C.; Raja Mahmood, R.A.; Adam, M.S.; Bakar, U.A.; Abdullah, N.F.; Luong, N.C. A Comprehensive Survey of Deep Reinforcement Learning in UAV-Assisted IoT Data Collection. *Veh. Commun.* **2025**, *55*, 100949. [\[CrossRef\]](#)
27. Shen, Y.; Xie, L.; Li, M. Intelligent Path Selection Algorithm for Tactical Communication Networks Enhanced by Link State Awareness. *Front. Commun. Netw.* **2025**, *6*, 1635982. [\[CrossRef\]](#)
28. Zhang, R.; Xiong, K.; Lu, Y.; Fan, P.; Ng, D.W.K.; Letaief, K.B. Energy Efficiency Maximization in RIS-Assisted SWIPT Networks With RSMA: A PPO-Based Approach. *IEEE J. Sel. Areas Commun.* **2023**, *41*, 1413–1430. [\[CrossRef\]](#)
29. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* **2021**, *372*, n71. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Viana, J.; Farkhari, H.; Gil Jiménez, V.P. Securing 5G and Beyond-Enabled UAV Links: Resilience Through Multiagent Learning and Transformers Detection. *IEEE Access* **2025**, *13*, 153993–154007. [\[CrossRef\]](#)
31. Zuo, P.; Miao, C.; Fu, C.; Wang, X.; Liu, X.; Liu, B. SMAPPO: A Security-Aware Multi-Agent Reinforcement Learning Framework for Secure Computation Offloading in SAGIN. *J. King Saud Univ. Comput. Inf. Sci.* **2025**, *37*, 336. [\[CrossRef\]](#)
32. Ericsson. Growth of Mobile Network Data Traffic Persists. Available online: <https://www.ericsson.com/en/reports-and-papers/mobility-report/dataforecasts/mobile-traffic-forecast> (accessed on 24 June 2026).

33. Bokobza, Y.; Dabora, R.; Cohen, K. Deep Reinforcement Learning for Simultaneous Sensing and Channel Access in Cognitive Networks. *IEEE Trans. Wirel. Commun.* **2023**, *22*, 4930–4946. [[CrossRef](#)]
34. Bai, W.; Zheng, G.; Xia, W.; Mu, Y.; Xue, Y. Multi-User Opportunistic Spectrum Access for Cognitive Radio Networks Based on Multi-Head Self-Attention and Multi-Agent Deep Reinforcement Learning. *Sensors* **2025**, *25*, 2025. [[CrossRef](#)] [[PubMed](#)]
35. Hong, S.; Jeong, Y.; Hwang, U.; Hong, S. QIPPO/CA: A Quantized Communication-Efficient MARL Framework for Fully Distributed Channel Access in Next-Generation Wireless Networks. *IEEE Internet Things J.* **2026**, *13*, 8615–8627. [[CrossRef](#)]
36. Wang, Q.; Xu, W.; Chen, H.-H. A Heterogeneous-Agent Deep Reinforcement Learning Approach for Dynamic Spectrum Access in Cognitive Wireless Networks. *IEEE Trans. Cogn. Commun. Netw.* **2025**, *12*, 2221–2235. [[CrossRef](#)]
37. Wang, T.; Cheng, K.; Du, X. Multi-Agent Independent PPO-Based Automatic ECN Tuning for High-Speed Data Center Networks. In Proceedings of the 2025 IEEE International Conference on Cluster Computing (CLUSTER), Edinburgh, UK, 2–5 September 2025; pp. 1–11.
38. Lin, G.; Xiao, Y.; Yu, S.; Yu, K.; Liu, J. Constrained Probabilistic Routing with RouterRL: A General Packet-Level Network Simulation Framework. *IEEE Trans. Netw. Sci. Eng.* **2026**, *13*, 2604–2622. [[CrossRef](#)]
39. Sefati, S.S.; Haq, A.U.; Nidhi; Craciunescu, R.; Halunga, S.; Mihovska, A.; Fratu, O. A Comprehensive Survey on Resource Management in 6G Network Based on Internet of Things. *IEEE Access* **2024**, *12*, 113741–113784. [[CrossRef](#)]
40. Xu, Y.; Zhu, K.; Xu, H.; Ji, J. Deep Reinforcement Learning for Multi-Objective Resource Allocation in Multi-Platoon Cooperative Vehicular Networks. *IEEE Trans. Wirel. Commun.* **2023**, *22*, 6185–6198. [[CrossRef](#)]
41. Alsahfi, T.; Badshah, A.; Alsini, R.; Shoi Alallah, F.; Bedewi, W.; Daud, A. Proximal Policy Optimization for Vehicular Big Data Offloading Across Edge, Regional, and Cloud Layers. *J. Grid Comput.* **2025**, *23*, 28. [[CrossRef](#)]
42. Rehman, A.U.; Sualiheen, S.; Chang, K. Adaptive Resource Allocation in 6G A2G-TN Integrated System: A Cross-Layer Multi-Agent PPO Approach. *Comput. Netw.* **2025**, *272*, 111655. [[CrossRef](#)]
43. Shang, C.; Yu, J.; Thai Hoang, D. Energy-Efficient and Intelligent ISAC in V2X Networks with Spiking Neural Networks-Driven DRL. *IEEE Trans. Wirel. Commun.* **2026**, *25*, 1182–1195. [[CrossRef](#)]
44. Kahraman, İ.; Köse, A.; Koca, M.; Anarim, E. Age of Information in Internet of Things: A Survey. *IEEE Internet Things J.* **2024**, *11*, 9896–9914. [[CrossRef](#)]
45. Wang, W.; Chen, Q.; Shen, Y.; Xiang, Z. Leakage Identification of Underground Structures Using Classification Deep Neural Networks and Transfer Learning. *Sensors* **2024**, *24*, 5569. [[CrossRef](#)] [[PubMed](#)]
46. Ayyappan, V.; Bruno, M.A. Applying Machine Learning to Optimize Resource Allocation & Maritime Wireless Mobile Network. *J. Wirel. Mob. Netw. Ubiquitous Comput. Dependable Appl.* **2025**, *16*, 406–416. [[CrossRef](#)]
47. Lin, X.; Liu, A.; Han, C.; Liang, X.; Sun, Y.; Ding, G.; Zhou, H. Intelligent Adaptive MIMO Transmission for Nonstationary Communication Environment: A Deep Reinforcement Learning Approach. *IEEE Trans. Commun.* **2025**, *73*, 5965–5979. [[CrossRef](#)]
48. Khaskheli, M.B.; Zhao, Y.; Lai, Z. Sustainable Maritime Governance of Digital Technologies for Marine Economic Development and for Managing Challenges in Shipping Risk: Legal Policy and Marine Environmental Management. *Sustainability* **2025**, *17*, 9526. [[CrossRef](#)]
49. Mao, Z.; Zhang, Z.; Lu, F.; Liu, X.; Xu, Z.; Pan, Y.; Kang, J.; You, Y. Dynamic Joint Resource Allocation in Maritime Wireless Communication Networks: A Meta-Reinforcement Learning Approach Based on Knowledge Embedding. *Front. Inf. Technol. Electron. Eng.* **2025**, *26*, 2672–2687. [[CrossRef](#)]
50. Lin, T.; Du, J.; Zhang, H.; Nallanathan, A.; Wang, J. PPO-Based Energy-Efficient Power Control and Spectrum Allocation in In-Vehicle HetNets. In Proceedings of the GLOBECOM 2023—2023 IEEE Global Communications Conference, Kuala Lumpur, Malaysia, 4–8 December 2023; pp. 6334–6339.
51. Ma, W.; Lin, B.; Pan, H.; Sun, G.; Shi, E.; An, J.; Yuen, C. SIM-Assisted Secure Mobile Communications via Enhanced Proximal Policy Optimization Algorithm. *IEEE Trans. Wirel. Commun.* **2026**, *25*, 11964–11979. [[CrossRef](#)]
52. Fang, J.; Wang, X.; Liu, Y.; Tang, H.; Li, X. Multi-Agent Collaborative Inference Optimization for Large-Scale DNNs in IoT Edge Systems. *IEEE Internet Things J.* **2026**, *13*, 24938–24953. [[CrossRef](#)]
53. Ferrag, M.A.; Friha, O.; Kantarci, B.; Tihanyi, N.; Cordeiro, L.; Debbah, M.; Hamouda, D.; Al-Hawawreh, M.; Choo, K.-K.R. Edge Learning for 6G-Enabled Internet of Things: A Comprehensive Survey of Vulnerabilities, Datasets, and Defenses. *IEEE Commun. Surv. Tutor.* **2023**, *25*, 2654–2713. [[CrossRef](#)]
54. Mustafa, E.; Shuja, J.; Rehman, F.; Namoun, A.; Bilal, M.; Iqbal, A. Computation Offloading in Vehicular Communications Using PPO-Based Deep Reinforcement Learning. *J. Supercomput.* **2025**, *81*, 547. [[CrossRef](#)]
55. Soni, L.; Taneja, A.; Alqahtani, N.; Alqahtani, J. Robust ISAC Based Framework for Location Estimation and Target Detection in 6G Networks. *PLoS ONE* **2026**, *21*, e0337050. [[CrossRef](#)] [[PubMed](#)]
56. Jabeen, N.; Lei, H.; Muhammad, A.; Ali, A.; Khan, Z.U.; Pan, G. Localization in ISAC: A Review. *IEEE Internet Things J.* **2025**, *12*, 46526–46552. [[CrossRef](#)]
57. Wu, K.; Wang, Z.; Chen, S.-L.; Zhang, J.A.; Guo, Y.J. ISAC: From Human to Environmental Sensing. *IEEE J. Sel. Top. Electromagn. Antennas Propag.* **2025**, *1*, 84–98. [[CrossRef](#)]

58. Wang, K.; Sun, Y.; Liu, P.; Zhang, Y.; Shao, Z. Energy-Efficient Deep Reinforcement Learning RSMA in Multi-UAV-Assisted Wireless-Powered Communication Network. *IEEE Trans. Netw. Sci. Eng.* **2026**, *13*, 2420–2438. [\[CrossRef\]](#)
59. Guan, W.; Zhang, H. Introduction. In *Network Slicing for Future Wireless Communication; Wireless Networks*; Springer Nature: Cham, Switzerland, 2024; pp. 1–12.
60. Donatti, A.W.; Cristina Machado, M.; Alexander Lopez Martinez, M.; Rogério Antunes, S.S.; Carlos Figueiredo Souza, E.; Corrêa, S.L.; Ferreto, T.C.; Augusto Suruagy, J.; Martins, J.S.B.; Cristina Carvalho, T. Energy Efficiency in Network Slicing: Survey and Taxonomy. *IEEE Access* **2025**, *13*, 134570–134589. [\[CrossRef\]](#)
61. Dubey, M.; Singh, A.K.; Mishra, R. AI Based Resource Management for 5G Network Slicing: History, Use Cases, and Research Directions. *Concurr. Comput. Pract. Exp.* **2025**, *37*, e8327. [\[CrossRef\]](#)
62. Mahmood, A.; Abdallah, A.M.; Baharom, B.B.; Habilah, A.S.K. Ensuring Satellite Operational Integrity: A Power Budget Analysis for Next Generation Satellites. *IEEE Access* **2025**, *13*, 44901–44911. [\[CrossRef\]](#)
63. Chen, R.; Long, W.-X.; Wang, B.; He, Y.; Sun, R.; Cheng, N.; Zheng, G.; Niyato, D. Multibeam High Throughput Satellite: Hardware Foundation, Resource Allocation, and Precoding. *IEEE Commun. Surv. Tutor.* **2026**, *28*, 5379–5415. [\[CrossRef\]](#)
64. Zhao, D.; Wang, Y.; Song, B.; Zhou, Y.; Qin, P. Learning When and Where to Handover: A Hierarchical Reinforcement Learning Framework for Dense LEO Satellite Constellations. *IEEE Trans. Wirel. Commun.* **2026**, *25*, 12787–12801. [\[CrossRef\]](#)
65. Marzuk, F.; Vejar, A.; Cholda, P. Deep Reinforcement Learning for Energy-Efficient 6G V2X Networks. *Electronics* **2025**, *14*, 1148. [\[CrossRef\]](#)
66. Li, Y.; Chang, Y.; Fukawa, K.; Kodama, N. Reinforcement Learning-Based Cognitive Radio Transmission Scheduling in Vehicular Systems. In Proceedings of the 2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring), Florence, Italy, 20–23 June 2023; pp. 1–5.
67. Su, W.; Liu, H.; Li, T.; Lv, X.; Rui, H.; Huang, W.; Wang, Z.; Li, Y. Jointly Optimizing Deployment and Antenna of Base Stations Using Hierarchical Reinforcement Learning. *ACM Trans. Knowl. Discov. Data* **2026**, *20*, 1–25. [\[CrossRef\]](#)
68. Sun, K.; Yang, J.; Li, J.; Yang, B.; Ding, S. Proximal Policy Optimization-Based Hierarchical Decision-Making Mechanism for Resource Allocation Optimization in UAV Networks. *Electronics* **2025**, *14*, 747. [\[CrossRef\]](#)
69. Li, C.; Tan, X.; Chen, C. Deep Reinforcement Learning–Driven Multi-Satellite Collaborative Observation Planning for Emergency and Disaster Monitoring. *Int. J. Digit. Earth* **2025**, *18*, 2554310. [\[CrossRef\]](#)
70. He, Z.; Luo, Y.; Shojafar, M.; Mi, D. Heterogeneous-Agent PPO RL for xApps Coordination in Digital Twin Enabled O-RAN. In Proceedings of the 2025 IEEE/CIC International Conference on Communications in China (ICCC), Shanghai, China, 10–13 August 2025; pp. 1–6.
71. Qazzaz, M.M.H.; Salama, A.; Hafeez, M.; Zaidi, S.A.R. OREO: Open RAN Energy Optimization via Deep Reinforcement Learning for 6G Networks. *IEEE Open J. Commun. Soc.* **2026**, *7*, 4165–4182. [\[CrossRef\]](#)
72. Asemian, G.; Amini, M.; Kantarci, B. Anti-Jamming Task Scheduling in MEC-O-RAN With Hierarchical DRL and Transformer-Based Control. *IEEE Internet Things J.* **2026**, *13*, 7714–7729. [\[CrossRef\]](#)
73. Di, Z.; Zhong, Z.; Pengfei, Q.; Hao, Q.; Bin, S. Resource Allocation in Multi-User Cellular Networks: A Transformer-Based Deep Reinforcement Learning Approach. *China Commun.* **2024**, *21*, 77–96. [\[CrossRef\]](#)
74. Lee, T.H.; Kim, J. Hybrid PPO–DQN for Multi-Objective Adaptive Cruise Control in Eco-Driving: Reward Shaping Toward Safety and Sustainability (Student Abstract). In Proceedings of the AAAI Conference on Artificial Intelligence, Singapore, 20–27 January 2026.
75. Bereketeab, L.; Zekeria, A.; Aloqaily, M.; Guizani, M.; Debbah, M. Energy Optimization in Sustainable Smart Environments with Machine Learning and Advanced Communications. *IEEE Sens. J.* **2024**, *24*, 5704–5712. [\[CrossRef\]](#)
76. Song, J.; Gao, Y.; Wu, D.; Zhou, L. Adaptive Live Tactile Streaming with Scalable Coding for Immersive Communications. *IEEE Trans. Mob. Comput.* **2026**, *25*, 5133–5145. [\[CrossRef\]](#)
77. Jihad, M.; Al Fahad, A.; Roy, P.; Razzaque, M.A.; Alelaiwi, A.; Hassan, M.R.; Hassan, M.M. Quality of Experience Aware Task Execution in Digital Twinning Vehicular Edge Computing: A Framework and A3C Algorithm. *Future Gener. Comput. Syst.* **2026**, *176*, 108144. [\[CrossRef\]](#)
78. Luo, B.; Lu, X.; Lu, W.; Han, H.; Huang, G.; Zhang, Y. Neural Adaptive Video Streaming via Imitation Learning and Reinforcement Learning. In Proceedings of the 2024 IEEE 10th International Symposium on Microwave, Antenna, Propagation and EMC Technologies for Wireless Communications (MAPE), Guangzhou, China, 27–30 November 2024; pp. 1–4.
79. Iqbal, A.; Al-Habashna, A.; Wainer, G.; Boudreau, G. Sum Rate Maximization in RIS-Assisted Multi-User MISO Systems: A Proximal Policy Optimization-Based Approach. *Phys. Commun.* **2026**, *74*, 102961. [\[CrossRef\]](#)
80. Wara, N.; Paul, A.; Singh, K.; Kaushik, A.; Shin, W. Multi-Agent PPO-Based Resource Optimization for Full-Duplex RIS-Aided NOMA-ISAC Systems. *IEEE Open J. Commun. Soc.* **2025**, *6*, 9802–9820. [\[CrossRef\]](#)
81. Hu, J.; Ju, Y.; Wang, H.; Liu, L.; Pei, Q.; Guo, Y.; Wu, C. Multi-RIS Intelligent Collaboration Empowered Secure MmWave D2D Communication. In Proceedings of the GLOBECOM 2024—2024 IEEE Global Communications Conference, Cape Town, South Africa, 8–12 December 2024; pp. 3243–3248.

82. Iqbal, A.; Al-Habashna, A.; Wainer, G.; Boudreau, G.; Bouali, F. PPO-Based Energy Efficiency Maximization For RIS-Assisted Multi-User Miso Systems. In Proceedings of the 2024 IEEE 100th Vehicular Technology Conference (VTC2024-Fall), Washington, DC, USA, 7–10 October 2024; pp. 1–6.
83. Wu, Y.; Huang, Y.; Wang, Z.; Xu, C. Joint Caching and Computation in UAV-Assisted Vehicle Networks via Multi-Agent Deep Reinforcement Learning. *Drones* **2025**, *9*, 456. [\[CrossRef\]](#)
84. O'Connell, E.; O'Brien, W.; Bhattacharya, M.; Moore, D.; Penica, M. Digital Twins: Enabling Interoperability in Smart Manufacturing Networks. *Telecom* **2023**, *4*, 265–278. [\[CrossRef\]](#)
85. Yang, B.; Wu, B.; You, Y.; Guo, C.; Qiao, L.; Lv, Z. Edge Intelligence Based Digital Twins for Internet of Autonomous Unmanned Vehicles. *Softw. Pract. Exp.* **2024**, *54*, 1833–1851. [\[CrossRef\]](#)
86. Zhang, P.; Li, G. A Cooperative Dynamic Target Search Approach for Multi-UAV Systems Utilizing the MAPPO Algorithm. *Discov. Artif. Intell.* **2025**, *5*, 153. [\[CrossRef\]](#)
87. Xie, W.; Yang, H.; Xiong, Z. Resource Allocation for UAV-Assisted Anti-Jamming Semantic D2D Networks: A Graph Reinforcement Learning Approach. *Comput. Netw.* **2025**, *269*, 111463. [\[CrossRef\]](#)
88. He, H.; Yang, X.; Mi, X.; Shen, H.; Liao, X. Multi-Agent Deep Reinforcement Learning Based Dynamic Task Offloading in a Device-to-Device Mobile-Edge Computing Network to Minimize Average Task Delay with Deadline Constraints. *Sensors* **2024**, *24*, 5141. [\[CrossRef\]](#) [\[PubMed\]](#)
89. Hu, F.; Fu, Q.; Zhang, S.; Huang, J. A Multi-Agent Deep Reinforcement Learning-Based Task Offloading Method for 6G-Enabled Internet of Vehicles with Cloud-Edge-Device Collaboration. *Comput. Mater. Contin.* **2026**, *87*, 1. [\[CrossRef\]](#)
90. Hu, Z.; Liu, X.; Guo, M.; Liu, C. PPO-Based Joint Task Offloading and Resource Allocation for Vehicular Edge Computing Via V2I and V2V Communications. In Proceedings of the 2025 13th International Conference on Intelligent Computing and Wireless Optical Communications (ICWOC), Chengdu, China, 28–29 June 2025; pp. 316–321.
91. Singh, R.; Kaushik, A.; Shin, W.; Renzo, M.D.; Sciancalepore, V.; Lee, D.; Sasaki, H.; Shojaeifard, A.; Dobre, O.A. Toward 6G Evolution: Three Enhancements, Three Innovations, and Three Major Challenges. *IEEE Netw.* **2025**, *39*, 139–147. [\[CrossRef\]](#)
92. Mohammed, S.A.; Murad, S.S.; Albeyboni, H.J.; Soltani, M.D.; Ahmed, R.A.; Badeel, R.; Chen, P. Supporting Global Communications of 6G Networks Using AI, Digital Twin, Hybrid and Integrated Networks, and Cloud: Features, Challenges, and Recommendations. *Telecom* **2025**, *6*, 35. [\[CrossRef\]](#)
93. Luo, X.; Lin, Q.; Zhang, R.; Chen, H.-H.; Wang, X.; Huang, M. ISAC—A Survey on Its Layered Architecture, Technologies, Standardizations, Prototypes, and Testbeds. *IEEE Commun. Surv. Tutor.* **2026**, *28*, 485–526. [\[CrossRef\]](#)
94. Mata, L.; Sousa, M.; Vieira, P.; Queluz, M.P.; Rodrigues, A. Optimizing Energy and Spectral Efficiency in Mobile Networks: A Comprehensive Energy Sustainability Framework for Network Operators. *IEEE Access* **2025**, *13*, 22342–22364. [\[CrossRef\]](#)
95. Shen, L.; Li, B.; Zhu, X. Robust Offloading for Edge Computing-Assisted Sensing and Communication Systems: A Deep Reinforcement Learning Approach. *Sensors* **2024**, *24*, 2489. [\[CrossRef\]](#) [\[PubMed\]](#)
96. Sun, A.; Wu, P.; Jin, S.; Jiao, F. Deep Reinforcement Learning-Driven Multidomain Resource Allocation for Integrated Sensing, Communication, and Computing. In Proceedings of the Eighth International Conference on Artificial Intelligence and Pattern Recognition (AIPR 2025), Quanzhou, China, 19–21 December 2025; Tian, H., Ed.; SPIE: Bellingham, WA, USA, 2025; p. 239.
97. Zhou, Y.; Feng, Z.; Wei, Z.; Ma, D.; Huang, D.; Meng, Z.; Fan, Y.; Xu, J.; Zhang, P. Integrated Sensing, Communication, and Control for Multi-Agent Networked Formation Control. *Sci. China Inf. Sci.* **2026**, *69*, 142301. [\[CrossRef\]](#)
98. Kastwar, N.; Gupta, A. Drone Taxis and Drone Deliveries in Smart Logistics Security Surveillance: An Examination of the Drone Revolution from Legal and Ethical Dimensions. In *Innovative Strategies in Aviation Management and Marketing*; Sousa, B.B., Marques, M.I., Arantes, L., O'Neill, A., Eds.; IGI Global Scientific Publishing: Palmdale, PA, USA, 2025; pp. 149–180.
99. Manda, V.K.; Christy, V.; Hlali, A. Current Trends, Opportunities, and Futures Research Directions in Geospatial Technologies for Smart Cities. In *Advances in Geospatial Technologies*; Darwish, D., Chemingui, H., Eds.; IGI Global: Hershey, PA, USA, 2024; pp. 239–270.
100. Ghomri, B.I.-D.; Bendimerad, M.Y.; Bendimerad, F.T. DRL-Driven Optimization for Energy Efficiency and Fairness in NOMA-UAV Networks. *IEEE Commun. Lett.* **2024**, *28*, 1048–1052. [\[CrossRef\]](#)
101. Aung, P.S.; Nguyen, L.X.; Tun, Y.K.; Han, Z.; Hong, C.S. Aerial STAR-RIS Empowered MEC: A DRL Approach for Energy Minimization. *IEEE Wirel. Commun. Lett.* **2024**, *13*, 1409–1413. [\[CrossRef\]](#)
102. Wang, Y.; Cui, Y.; Yang, Y.; Li, Z.; Cui, X. Multi-UAV Path Planning for Air-Ground Relay Communication Based on Mix-Greedy MAPPO Algorithm. *Drones* **2024**, *8*, 706. [\[CrossRef\]](#)
103. Chaudhary, S.; Budhiraja, I.; Chaudhary, R.; Garg, S.; Choi, B.J.; Alrashoud, M. Proximal Policy Optimization Based Sum Rate Maximization Scheme for STAR-RIS-Assisted Vehicular Networks Underlying UAV. *Alex. Eng. J.* **2025**, *118*, 700–710. [\[CrossRef\]](#)
104. Zhou, W.; Yi, M.; Zhang, Y.; Wang, X.; Liu, J. Satellite-Assisted UAV Data Collection for Information Freshness in IoRT Networks. In Proceedings of the 2024 IEEE Wireless Communications and Networking Conference (WCNC), Dubai, United Arab Emirates, 21–24 April 2024; pp. 1–6.

105. Boufakhreddine, Z.; Nohra, A.; Haidar, G.A.; Achkar, R.; Owayjan, M. Exploring the Potential of AI in Network Slicing for 5G Networks: An Optimisation Framework. *IET Commun.* **2025**, *19*, e70116. [\[CrossRef\]](#)
106. Nguyen, T.T.H.; Kim, T. Proximal Policy Optimization for Up/Downlink Time Slots Allocation in 5/6G Dynamic TDD Networks. *KSII Trans. Internet Inf. Syst.* **2025**, *19*, 259–278. [\[CrossRef\]](#)
107. Adhikari, B.; Shaharyar Khwaja, A.; Jaseemuddin, M.; Anpalagan, A. DRL-Leveraged and RIS-Assisted Hybrid Network Slicing for eMBB and URLLC Co-Existence in 6G Systems. *IEEE Open J. Commun. Soc.* **2025**, *6*, 6156–6176. [\[CrossRef\]](#)
108. Raja, G.; Sanjeev, A.; Ravishankar, K.; Arunachalam, K. TRIP-B5G: Traffic Classification and Resource Allocation Using Intelligent PPO in B5G O-RAN. In Proceedings of the 2026 IEEE 23rd Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 9–12 January 2026; pp. 1–4.
109. Zhang, C.; Wu, C.; Lin, M.; Lin, Y.; Liu, W. Proximal Policy Optimization for Efficient D2D-Assisted Computation Offloading and Resource Allocation in Multi-Access Edge Computing. *Future Internet* **2024**, *16*, 19. [\[CrossRef\]](#)
110. Hikmat, F.A.; Sahib, M.A. PPO-Based Deep Reinforcement Learning Framework for Dynamic Resource Allocation and Network Slicing in 5G Mobile Networks. *Int. J. Electron. Telecommun.* **2026**, *72*, 1–9. [\[CrossRef\]](#)
111. Cui, Z.; Qamar, F.; Kazmi, S.H.A.; Zainol Ariffin, K.A.; Safdar, G.A.; Ur Rehman, M.H. A Review of Multi-Agent Deep Reinforcement Learning for Resource Allocation in beyond 5G Network Slicing: Solutions, Challenges and Future Research Directions. *PeerJ Comput. Sci.* **2026**, *12*, e3728. [\[CrossRef\]](#)
112. Yang, L.; Bi, Z.; Wang, Z.; Liang, X.; Zhang, J.; Wu, R. Resource Allocation for SFC Networks: A Deep Reinforcement Learning Approach. In Proceedings of the 2024 7th World Conference on Computing and Communication Technologies (WCCCT), Chengdu, China, 12–14 April 2024; pp. 210–215.
113. Zhang, Y.; Joe, I. The Optimized Deployment of Service Function Chain Based on Deep Reinforcement Learning Algorithm. In *Proceedings of the International Conference on Machine Learning, Pattern Recognition and Automation Engineering, Singapore, 7–9 August 2024*; ACM: New York, NY, USA, 2024; pp. 24–28.
114. Fu, S.; Wei, W.; Feng, X.; Yin, L. Average Sum Rate Optimization in RIS-Assisted NOMA Satellite Network: A Deep Reinforcement Learning Approach. *IEEE Wirel. Commun. Lett.* **2025**, *14*, 1772–1776. [\[CrossRef\]](#)
115. Zhang, H.; Han, X.; Xing, C.; Chen, H.; Zhao, J. Analysis of Uplink Transmission Scheduling Strategies for LoRa-Based Direct-to-Satellite IoT Networks Using Deep Reinforcement Learning. *IEEE Trans. Green Commun. Netw.* **2026**, *10*, 1279–1292. [\[CrossRef\]](#)
116. Xu, J.; Zhao, Z.; Wang, L.; Zhang, Y. A Novel Deep Reinforcement Learning Architecture for Dynamic Power and Bandwidth Allocation in Multibeam Satellites. *Acta Astronaut.* **2023**, *204*, 73–82. [\[CrossRef\]](#)
117. Li, S.; Wu, Q.; Wang, R. Efficient Packet Routing in Ultra-Dense LEO Satellite Networks via Cooperative-MARL with Queuing Theory Model. In Proceedings of the 2025 IEEE Wireless Communications and Networking Conference (WCNC), Milan, Italy, 24–27 March 2025; pp. 1–6.
118. Zhang, Q.; Fu, S.; Yang, Z. Jointly Optimizing Satellite Handover and Power Allocation in LEO Satellite Network: A Dual-Agent Framework. *IEEE Trans. Veh. Technol.* **2026**, 1–6. [\[CrossRef\]](#)
119. Sadiki, A.; Bentahar, J.; Dssouli, R.; En-Nouaary, A.; Otrouk, H. Deep Reinforcement Learning for the Computation Offloading in MIMO-Based Edge Computing. *Ad Hoc Netw.* **2023**, *141*, 103080. [\[CrossRef\]](#)
120. Li, Z.; Tian, J.; Zhang, H.; Shi, T.; Xu, B.; Zhou, T. A Multi-Agent Proximal Policy Optimization-Based Handover Scheme for Satellite-Terrestrial Integrated Networks. *IEEE Wirel. Commun. Lett.* **2026**, *15*, 2428–2432. [\[CrossRef\]](#)
121. Zhao, R.; Cai, J.; Luo, J.; Ran, Y.; Gao, J.; Xu, Y. Joint Beam Hopping and Resource Allocation for Load Balancing and Interference Avoidance in Multi-LEO Satellite Networks. In Proceedings of the ICC 2025—IEEE International Conference on Communications, Montreal, QC, Canada, 8–12 June 2025; pp. 958–963.
122. Meng, M.; Hu, B.; Chen, S.; Kang, S. Joint Beamforming and Dynamic Beam Hopping Based on MAPPO for LEO Satellite Communication System. *IEEE Wirel. Commun. Lett.* **2025**, *14*, 1461–1465. [\[CrossRef\]](#)
123. Alharbi, S. A Review of Deep Multi-Objective Reinforcement Learning and Vision-Based Systems for Smart Cities. *Informatica* **2025**, *49*. [\[CrossRef\]](#)
124. Louati, A.; Louati, H.; Kariri, E.; Neifar, W.; Hassan, M.K.; Khairi, M.H.H.; Farahat, M.A.; El-Hoseny, H.M. Sustainable Smart Cities through Multi-Agent Reinforcement Learning-Based Cooperative Autonomous Vehicles. *Sustainability* **2024**, *16*, 1779. [\[CrossRef\]](#)
125. Wang, J.; Wang, R.; Zheng, Z.; Lin, R.; Wu, L.; Shu, F. Physical Layer Security Enhancement in AAV-Assisted Cooperative Jamming for Cognitive Radio Networks: A MAPPO-LSTM Deep Reinforcement Learning Approach. *IEEE Trans. Veh. Technol.* **2025**, *74*, 4713–4727. [\[CrossRef\]](#)
126. Chen, M.; Chen, X.; Wang, R.; Ding, H. Reactive Jamming Resilient Power Allocation in Cognitive Radio Networks via Deep Reinforcement Learning. In *Intelligent Networked Things*; Zhang, L., Yu, W., Laili, Y., Qu, T., Eds.; Communications in Computer and Information Science; Springer Nature: Singapore, 2026; Volume 2624, pp. 327–335.
127. Zhang, F.; Niu, Y.; Zhou, W. Intelligent Anti-Jamming Decision Algorithm for Wireless Communication Based on MAPPO. *Electronics* **2025**, *14*, 462. [\[CrossRef\]](#)

128. Ma, H.; You, J.; Wu, H.; Xing, L.; Zhang, X. A Probabilistic Routing Algorithm Based on CNN and Q-Learning for Vehicular Edge Network. *Trans. Emerg. Telecommun. Technol.* **2025**, *36*, e70050. [\[CrossRef\]](#)
129. Alvarado-Padilla, J.J.; Celaya-Padilla, J.M.; Martinez-Torteya, A.; Soto-Murillo, M.A.; Gamboa-Rosales, H.; Gamboa-Rosales, N.K. Optimizing Autonomous Vehicle Control Through Deep Q-Learning: A Simulation-Based Approach with CARLA. In *Advanced Research in Technologies, Information, Innovation and Sustainability*; Guarda, T., Portela, F., Augusto, M.F., Eds.; Communications in Computer and Information Science; Springer Nature: Cham, Switzerland, 2025; Volume 2348, pp. 141–154.
130. Pimenow, S.; Pimenowa, O.; Prus, P. Challenges of Artificial Intelligence Development in the Context of Energy Consumption and Impact on Climate Change. *Energies* **2024**, *17*, 5965. [\[CrossRef\]](#)
131. Samaniego, J.F. From Hyperconnectivity to AI: How Can We Tackle the Environmental Impact of the New Internet Era? Available online: <https://www.uoc.edu/en/news/2023/080-sustainable-digitalization> (accessed on 24 June 2026).
132. Zhen, Y.; Tao, L.; Wu, D.; Tang, T.; Wang, R. Energy-Saving Control Strategy for Ultra-Dense Network Base Stations Based on Multi-Agent Reinforcement Learning. *Digit. Commun. Netw.* **2025**, *11*, 1007–1017. [\[CrossRef\]](#)
133. Wang, J.; Li, Y.; Hong, Y.; Tang, Y. Integrated Adaptive Communication in Multi-Agent Systems: Dynamic Topology, Frequency, and Content Optimization for Efficient Collaboration. *Neurocomputing* **2025**, *617*, 129068. [\[CrossRef\]](#)
134. Li, S.; Fang, B. AI Development Strategies in Countries Around the World. In *Artificial Intelligence Security and Safety*; Fang, B., Ed.; Springer Nature: Singapore, 2025; pp. 51–84.
135. García-Pineda, V.; Valencia-Arias, A.; Patiño-Vanegas, J.C.; Flores Cueto, J.J.; Arango-Botero, D.; Rojas Coronel, A.M.; Rodríguez-Correa, P.A. Research Trends in the Use of Machine Learning Applied in Mobile Networks: A Bibliometric Approach and Research Agenda. *Informatics* **2023**, *10*, 73. [\[CrossRef\]](#)
136. Espinosa, A.; Samos, X.; Ulied, D.; Marias, J.; Touma, R. Optimizing Energy Consumption of Edge-Cloud Environments: A Comparative Study Between PPO and PSO. *Int. J. Comput. Intell. Syst.* **2025**, *19*, 16. [\[CrossRef\]](#)
137. Wu, Z.; Fang, H.; Tang, J.; Yang, X. Lightweight Adaptive PPO-AHP Enhanced Algorithm for Task Offloading in Vehicular Edge Computing. In Proceedings of the 2025 International Joint Conference on Neural Networks (IJCNN), Rome, Italy, 30 June–5 July 2025; pp. 1–9.
138. Li, S.; Huang, Q.; Chen, H.; Jiang, R.; Dong, M.; Ota, K.; Quek, T.Q.S. DRL-Based Joint Task Offloading and Resource Allocation in Air-Ground Integrated Vehicular Edge Computing Network with Energy Harvesting. *IEEE Trans. Veh. Technol.* **2025**, *75*, 6658–6672. [\[CrossRef\]](#)
139. Wu, C.-M.; Guan, S.-Z.; Yang, C.-C.; Lin, K.-T.; Kuang, M.-Y. Scalable and Interference-Aware Spectrum Access in IoT Networks via Decentralized Proximal Policy Optimization. *Comput. Netw.* **2026**, *275*, 111885. [\[CrossRef\]](#)
140. Nzewi, O.I. Adaptive Governance for Resilient Local Service Delivery. *J. Local Gov. Res. Innov.* **2025**, *6*, a322. [\[CrossRef\]](#)
141. Hu, Y.; Cong, R.; Matsumoto, T.; Li, Y. Environmental and Economic Impacts of V2X Applications in Electric Vehicles: A Long-Term Perspective for China. *Energies* **2025**, *18*, 3636. [\[CrossRef\]](#)
142. Liu, X.; Tan, J.; Ren, X.; Dai, H. Self-Attention Proximal Policy Optimization for Beam Tracking in mmWave Communications. *IEEE J. Sel. Areas Commun.* **2026**, *44*, 4552–4569. [\[CrossRef\]](#)
143. Liu, C.; Zhao, J.; Li, J.; Wang, D.; Yu, F.R. UAV Aided Integrated Sensing, Communication and Computing: Optimization via Federated Learning. *IEEE Trans. Veh. Technol.* **2025**, *75*, 6045–6058. [\[CrossRef\]](#)
144. Hu, H.; Tang, H.; Zhang, R.; Jiang, F.; Ding, Z.; Niyato, D. Joint Scheduling and Power Control in AoI-Oriented WP-IoT Networks: An HVF-Based PPO Approach. *IEEE Trans. Veh. Technol.* **2025**, *75*, 6876–6881. [\[CrossRef\]](#)
145. Chen, L.; Wu, W.; Tian, F. Efficient Resource Allocation for RSMA-Based Semantic Image Transmission with Shared Knowledge Base. *IEEE Wirel. Commun. Lett.* **2026**, *15*, 2154–2158. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.