A Unified Taxonomy and Empirical Review of Recent Proximal Policy Optimization Variants and Their Real-World Applications

Vijaya Kittu Manda, Bhukya Madhu
BEST Innovation University, Gorantla, Andhra Pradesh, India
Malla Reddy (MR) Deemed to be University, Hyderabad, Telangana, India
E-mail: madhu0525@gmail.com, https://mrdu.edu.in/

Keywords: generalization, computational efficiency, robotic control, autonomous systems, route optimization

Received: October 25, 2025

Artificial Intelligence (AI) algorithms such as Proximal Policy Optimization (PPO) help train agents for sequential decision-making tasks. Existing surveys already provide good coverage of the original PPO and its early developments but fell short of capturing the rapid evolution of PPO-based methods over the recent few years. Since 2023, a wave of algorithmic variants has emerged. These variants address different objectives, regularization approaches, exploration methods, training pipelines, and hybrid architectures across diverse applications. However, there has been no systematic effort to organize, compare, or critically assess these advances. This review addresses that research gap. This review analyzes 32 peer-reviewed studies (2023–2025) to evaluate 15+ PPO variants across six innovation categories and ten application domains. The study revealed that PPO delivers a typical performance improvement of 15–44% over baselines across metrics, including improved safety constraint satisfaction (+15%), computational efficiency (+18% SLA compliance), and sim-to-real transfer (+23% task success). The study analyzed advancements and developments by proposing a unified taxonomy focused on algorithmic advances and their performance in real-world scenarios. Three critical dimensions considered for evaluation are: generalization across tasks and environments, robustness and safety in deployment, and computational efficiency in training and inference. The review also identifies recurring limitations, inconsistent evaluation practices, and underexplored directions. It exposes gaps between simulation benchmarks and real-world deployment conditions, including operational constraints and challenges. By connecting theoretical improvements to empirical outcomes, this work serves as both a practical reference for engineers and researchers applying PPO today. The synthesized taxonomy provides a structured reference for analyzing recent PPO variants and their empirical trade-offs.

Povzetek: Pregled analizira nove različice algoritma PPO (2023–2025), pokaže njihove izboljšave zmogljivosti (15–44 %) ter izpostavi ključne prednosti, omejitve in vrzeli med simulacijami in uporabo v praksi.

1 Introduction

Artificial Intelligence (AI) and Machine Learning (ML) have revolutionized complex decision-making tasks across various domains [1], [2]. Reinforcement Learning (RL) is a core subset of AI. It enables agents to learn optimal behaviors through trial-and-error interactions with dynamic environments [3]. Proximal Policy Optimization (PPO) has emerged as a leading framework for achieving stable, efficient policy improvement among RL methods [4]. PPO is flexible and robust. So, it is ideal for widespread adoption of robotics [5], autonomous systems [6], and dynamic resource management [7].

Previous surveys have laid essential groundwork by explaining PPO's foundational mechanics and early applications [8], [9]. However, they could not capture the rapid methodological evolution since 2023. Lately, over a dozen PPO variants have emerged. Significant architectural, safety-aware, and efficiency-oriented innovations have been made. Crucially, no existing survey offers a structured taxonomy grounded in algorithmic novelty, nor do they systematically benchmark performance trade-offs in real-world settings. Unlike previous PPO surveys that organize methods by application domains, this work introduces an informatics-oriented taxonomy. This taxonomy classifies PPO variants based on their algorithmic transformations and evaluates them using formally defined dimensions: generalization, safety, and computational efficiency.

The review addresses this gap through a systematic analysis of 32 peer-reviewed studies (2023–2025) that evaluate 15+ PPO variants across six innovation categories. It links theoretical modifications to empirical outcomes across various application areas, including robotics, autonomous systems, and constrained optimization. The review aligns with recent methodological developments in PPO research. PPO delivers stability and straightforward implementation through its core design. Researchers have since developed

numerous algorithmic innovations to address its limitations in safety, sample efficiency, and multi-agent coordination. Below, we categorize these advances through a taxonomy grounded in their primary technical contributions.

1.1 Key concepts of PPO

Proximal Policy Optimization (PPO) is a reinforcement learning algorithm developed by OpenAI in 2017 that has gained prominence for its practical stability and ease of implementation. It uses a trust-region mechanism that enables consistent policy improvement while preventing destabilizing updates that compromise training stability. The PPO algorithm extends the policy gradient and actor–critic lineages of reinforcement learning by simplifying the constrained update mechanism of Trust Region Policy Optimization (TRPO). While TRPO enforces strict trust regions via KL-divergence constraints that require second-order optimization, PPO achieves comparable stability. This is accomplished through a clipped surrogate objective that is open to first-order gradient methods [2]. This design explicitly constrains the magnitude of policy updates. It yields more reliable learning than Advantage Actor–Critic (A2C/A3C) variants. The variants lack explicit trust-region control, despite their synchronous or asynchronous update schemes [6].

1.2 Stability and efficiency with PPO

Naïve RL algorithms often exhibit training instability; consequently, numerous algorithmic variants (DQN, DDQN, SAC, DDPG, and PPO) are developed. PPO and related trust region methods provide stability guarantees by constraining policy updates to bounded regions, ensuring monotonic improvement and preventing destructive parameter shifts during training. REINFORCE is the foundational and first policy gradient model. Actor-Critic, TRPO, and PPO belong to this family of algorithms. PPO is a model-free, on-policy deep reinforcement learning (DRL) algorithm introduced by OpenAI in 2017. Because it updates using the most recent policy data, it provides more stable training than off-policy alternatives, such as the Soft Actor-Critic (SAC) algorithm. It also consistently outperforms SAC in managing workloads, serving more requests with less variance across training seeds. It also performs well at load balancing [8] and achieves lower rejection rates than similar algorithms [10]. PPO showed superior overall performance to other DRL algorithms (DQN, DDQN, DDPG). PPO achieved the highest cost reduction (12.3%), best schedule adherence (SPI: 1.18), and significant safety improvement (45%) [11]. It also showed a 44% improvement over traditional First-In-First-Out (FIFO) scheduling approaches in performance [11]. Policy gradient methods directly optimize parameterized policies through gradient-based updates, enabling stable learning in continuous and stochastic action spaces.

With this, the agents can learn stochastic policies, handle continuous action spaces, and make policy gradients. The tool is especially useful in complex, dynamic environments. These characteristics make policy gradient methods well-suited for real-world control tasks. These characteristics informed PPO's design philosophy of balancing policy improvement with update stability.

The PPO method utilizes Generalized Advantage Estimation (GAE) to compute the advantage function, thereby balancing bias and variation in policy evaluation [12]. The Advantage Function critically shapes PPO's performance by quantifying how much better or worse each action performs relative to the average actions available in a given state. The final optimization objective of the actor network in PPO is formulated to balance the new policy's performance and its deviation from the old policy.

This is done using a clipped surrogate objective function to ensure stable policy updates. The PPO method aims to improve the stability and efficiency of policy optimization. This is achieved by carefully controlling the extent of policy updates, thereby avoiding significant, drastic, or destructive policy changes that could lead to instability [13]. It introduces a soft trust region by clipping the probability ratio to stabilize training.

PPO improves Policy Gradient (PG) methods by addressing issues such as policy overfitting and suboptimal update directions resulting from variable step sizes. Using the PG method solves stochastic and dynamic challenges. It builds on Trust Region Policy Optimization (TRPO) and simplifies the trust-region constraints using Clipped Importance Sampling. This makes it more computationally efficient and easier to implement [14]. While prior surveys have covered foundational policy gradient methods or early PPO applications, this work provides a comprehensive analysis of the 2025 surge in PPO extensions. It systematically maps their performance trade-offs across real-world domains. We further identify a critical disconnect between simulation-based benchmarks and real-world deployment challenges. With this, the study offers a structured agenda for bridging this gap.

1.3 Improving computational efficiency

DRL models require significant computational resources primarily because they must repeatedly train and evaluate the model. However, across various RL models, PPO has an edge because it performs better computationally than alternatives such as TRPO. After all, significant policy updates can degrade performance. PPO is sample-efficient and computationally scalable, but not always computationally efficient. When dealing with millions of data points, PPO requires high-performance GPUs and cloud computing resources for training [15]. This computational demand can limit applicability in resource-constrained edge environments unless optimization techniques are employed. Hence, it is common for practitioners to optimize in the best possible way. This includes data preprocessing, improvements to the simulation environment, and hardware optimization to reduce the burden. Several strategies focus on model compression and architectural simplification to mitigate these hardware demands. The reported efficiency improvements of approximately 15–20% match ranges

commonly observed in prior PPO studies. Some studies opted for lightweight neural networks (CNN + Multi-Layer Perceptron (MLP) for real-time execution. A feedforward artificial neural network processes non-spatial inputs (e.g., agent position, global features). Beyond model design, algorithmic refinements improve training stability and efficiency, particularly during hyperparameter tuning. Nevertheless, despite these optimizations, several fundamental limitations persist in current PPO-based approaches.

1.4 Research objectives

The study aims to systematically classify PPO variants proposed since 2023 into innovation-driven taxonomic categories. It synthesizes their performance by gathering empirical evidence across generalization, safety, and computational efficiency dimensions.

1.5 Roadmap of this study

1.6 Review methodology

Section 1 formally introduces the research problem, the need for the study, and the novelty that the study will bring. Section 2, titled "Methods & Materials," details the literature review methodology and approach to conducting the review. Section 3 presents a novel taxonomy of PPO variants, while Section 4 surveys key application domains. Section 5 discusses the positioning of PPO, while Section 6 critically analyzes the limitations of existing research. Section 7 outlines promising future research directions. The conclusion remarks follow. Given PPO's foundational role in modern reinforcement learning, recent scholarly efforts have sought to refine, extend, and apply it across diverse domains. It motivated a focused review of the current literature. Following the establishment of PPO's foundational strengths and limitations, the methodology for curating and evaluating recent advances is detailed below.

2 Methods

This section discusses the method used to conduct the review and construct the taxonomy.

2.1 Review methodology

This review synthesizes peer-reviewed advances in Proximal Policy Optimization (PPO) variants published between 2023 and 2025. We prioritized publications that introduced novel algorithmic innovations. These included modified update mechanisms, safety constraints, or hybrid architectures, rather than straightforward applications of vanilla PPO. The taxonomy presented in Section 3 represents a conceptual synthesis of these studies, with variant definitions and application mappings extracted directly from published literature.

2.2 Review protocol and taxonomy construction procedure

The review is conducted using a systematic review protocol with the following explicit steps:

a. **Databases searched**: Scopus, IEEE Xplore, ACM Digital Library, and Springer Link.
b. **Time window**: Publications from January 2023 through December 2025.
c. **Inclusion/exclusion criteria**:
   i. Included: Peer-reviewed journal articles (Q1–Q2 Scopus quartiles) and rigorously reviewed conference proceedings (e.g., IEEE flagship conferences).
   ii. Excluded: Preprints (arXiv, ResearchGate), workshop papers, non-reviewed conference submissions, blog posts, and industry white papers.
   iii. Prioritized: Papers introducing novel PPO variants, fundamental theoretical modifications, or sophisticated hybrid architectures (e.g., PPO integrated with Transformers or Graph Neural Networks).
   iv. Deprioritized: Simple applications of standard PPO without algorithmic innovation (e.g., basic portfolio optimization or inventory management).
d. **Screening stages**:
   i. Stage 1: Title screening for relevance to PPO algorithmic variants.
   ii. Stage 2: Abstract screening for evidence of methodological novelty.
   iii. Stage 3: Full-text assessment for empirical validation and reproducibility.
e. **Data extraction fields**:
   i. Variant name and publication year
   ii. Core algorithmic innovation
   iii. Evaluation environments/datasets
   iv. Performance metrics and baseline comparisons
   v. Reported limitations
f. **Taxonomy assignment criteria**: Variants were categorized based on their primary technical contribution rather than application domain. Each variant was assigned to a single category reflecting its most significant innovation (e.g., safety constraint handling took precedence over multi-agent coordination when both features were present). Category boundaries were validated through iterative consensus among the authors using explicit decision rules.

Non-peer-reviewed publications, including preprints on ResearchGate or on personal websites/blogs, were excluded from the study. Conference papers that are not peer-reviewed but include peer-reviewed conference papers, primarily from IEEE. Publications are prioritized using established Scopus CiteScore and quartile rankings, with Q1 journals given the highest weight, followed by flagship IEEE conference proceedings. While conference publications are generally considered part of grey literature, rigorously peer-reviewed conference papers are

included in this category. It allows us to provide timely, rigorous dissemination of cutting-edge research. At the same time, priority is given to formal, peer-reviewed publications over preprints and reports from industry and universities. Rigorously peer-reviewed conference proceedings (e.g., IEEE flagship conferences) were included due to their timely dissemination of methodological innovations, whereas workshop papers and non-reviewed conference submissions were excluded. Preference was given to documents introducing a novel PPO variant, a fundamental theoretical modification, or sophisticated hybrid architecture (e.g., integrating PPO with Transformers or Graph Attention Networks). Simple applications of standard PPO in fields like finance or inventory management were deprioritized. This review emphasizes deployment-oriented PPO variants, prioritizing methods evaluated under operational constraints rather than purely benchmark-driven extensions. Regardless of venue quality, it is done to maintain the focus on core AI and algorithmic development. The taxonomy represents a conceptual synthesis of peer-reviewed PPO studies, with variant definitions, equations, and application mappings extracted directly from published literature.

Building on this work, researchers have proposed numerous algorithmic variants to address PPO's limitations and expand its capabilities.

2.3 Related work and comparative summary of PPO variants

Existing PPO surveys primarily focus on foundational mechanisms or early applications and do not systematically compare recent algorithmic variants emerging after 2023. While several studies report performance gains for individual PPO extensions, their evaluations are fragmented across heterogeneous benchmarks, metrics, and simulation settings. It made cross-variant comparison difficult. The fragmentation underscores the need for a unified comparative framework that synthesizes empirical evidence across variants and application domains. This framework directly informs the taxonomy proposed in our study.

Reported performance gains are drawn from original studies and reflect heterogeneous benchmarks, environments, and evaluation protocols. Values are presented to illustrate relative trends rather than strict head-to-head comparability. The comparative analysis in Table 1 reveals substantial diversity in evaluation practices, performance metrics, and validation depth across PPO variants. These inconsistencies obscure principled comparison and limit the transferability of reported gains. To address this gap, the following section introduces a unified taxonomy that organizes PPO variants by their core algorithmic innovations rather than by application-specific outcomes.

Table 1: Comparative Summary of Recent PPO Variants (2023–2025)

PPO Variant	Core Innovation	Dataset / Environment	Metrics Used	Reported Gain over Vanilla PPO	Key Limitations
PPO-Adaptive	Dynamically adjusts clipping range based on policy divergence	Smart grid simulators; cloud workload traces	Cumulative reward; convergence speed; SLA compliance	**+6–10%** faster convergence; **+8%** peak-load reduction	Increased per-iteration computation due to divergence monitoring; sensitive to threshold tuning
CPSPO	Couple penalty terms and clipping into a unified constrained objective	MuJoCo safety benchmarks; simulated healthcare control tasks	Reward, constraint violation rate, safety satisfaction	**+12–18%** higher reward under constraints; **30–40%** fewer violations	Requires careful penalty–clip balancing; limited evaluation beyond continuous control
MAPPO	Centralized critic with decentralized actors for multi-agent coordination	StarCraft II: multi-robot simulators	Win rate, coordination success, reward variance	**+15–25%** higher win rate vs. PPO; reduced variance	Centralized criticism limits scalability; communication overhead increases with the number of agents.
EMA-KPPO	K-hop localized information sharing in multi-agent PPO	UAV swarm simulations; decentralized energy systems	Communication rounds; task success; convergence rate	**~40%** reduction in communication; similar reward to MAPPO	Limited testing beyond structured graph environments; assumes known topology
Safe-PPO / PPO-Lagrangian	Lagrangian-based constraint enforcement	Autonomous driving simulators; robotic navigation	Constraint violations; episodic return	**20–35%** reduction in unsafe actions	Oscillatory constraint satisfaction; weak recovery analysis after violations
HPPO (JOR-HPPO)	Dual-actor architecture for discrete + continuous actions	Industrial IoT offloading; MEC simulations	Latency; energy cost; reward	**10–20%** latency reduction; improved feasibility	Architecture complexity; action-mask design is application-specific

PPO-DAPT	Domain-adaptive training for sim-to-real generalization	Robotics sim-to-real benchmarks	Task success under domain shift	**+20–25%** higher sim-to-real success	Requires domain randomization assumptions; limited real-world deployments
RPO (Reparameterized PPO)	Backpropagation through stochastic nodes for lower variance gradients	Continuous control (MuJoCo, PyBullet)	Sample efficiency; return	**~15%** fewer samples to reach the target reward	Not applicable to discrete or discontinuous action spaces
FedPPO	Federated PPO with decentralized data sharing	Distributed cloud/edge systems	SLA compliance; cost; convergence	**+15–18%** cost efficiency; high SLA adherence	Communication latency; non-IID data effects
Transformer-PPO	Transformer-based actor-critic for temporal dependencies	Robotics locomotion; sequential control tasks	Success rate, convergence steps	**+18–22%** task success; faster adaptation	High memory and compute cost; complex edge deployment
PPO-RND	Intrinsic motivation via Random Network Distillation	Sparse-reward games; exploration tasks	State coverage; reward	**Significant exploration gains** in sparse rewards	No benefit in dense-reward tasks; added network overhead
Quantum-Hybrid PPO	Hybrid quantum–classical policy/value networks	Small-scale control simulations	Training reward; stability	Higher training reward reported	Underperforms on real metrics; NISQ hardware limits scalability

3 Taxonomy of PPO Variants

3.1 PPO variants

This research theme focuses on methodological improvements, variations of the PPO algorithm, and new frameworks that aim to enhance the PPO optimization process or its performance. PPO is popular for handling continuous action spaces well and producing stochastic policies. This makes it suitable for applications that involve resource allocation [10], high-dimensional spaces [14], or multi-agent environments with partial observations [7]. Nevertheless, PPO variants can achieve more stable update mechanisms and reduced computational overhead compared to off-policy alternatives [16]. PPO gives better objective function values than Deep Deterministic Policy Gradient (DDPG). Among recently proposed variants, relatively few (CPSPO, EMA-KPPO, and PPO-DAPT) underwent rigorous validation across three or more distinct environments or included ablation studies. This shows a limited empirical validation for most proposals.

- **Collaborative Proximal Policy Optimization (CPPO)** employs a clipped expected discount reward mechanism to limit policy fluctuations during training. It demonstrates robust performance in safety-critical healthcare applications, such as personalized insulin dosing control.
- **Distributed Proximal Policy Optimization (DPPO)** is an RL model that optimizes decision-making in complex systems, such as power grids, by learning from environmental interactions [17].
- **Multi-Agent Proximal Policy Optimization (MAPPO)** uses a centralized critic with decentralized actors for multi-agent coordination [18]. The key variant, EMA-KPPO, introduces localized K-hop information sharing among agents, reducing communication rounds by ~40% while preserving convergence rate [19].
- **Coupled Penalties-Augmented Proximal Policy Optimization (CPSPO)**: CPSPO couples penalty terms and clipping into a unified constrained objective [20]. Unlike PPO-Lagrangian (oscillatory dynamics) and P3O (separate penalty/clipping), it prevents simultaneous violations of cost and policy constraints, outperforming mainstream safe RL algorithms in reward and constraint satisfaction on continuous control tasks.
- **Hybrid PPO (HPPO) framework** was developed to optimize discrete and continuous actions separately while accounting for interdependencies. In one such variant (JOR-HPPO), two parallel actor networks are used - one for discrete offloading decisions and another for continuous power allocation [14]. The application used a shared Critic network to guide policy updates. A beta distribution (instead of a Gaussian) was used to model continuous actions. It improved stability and compatibility with bounded variables, such as power transmission, and finally introduced an action masking mechanism. The mechanism filters out invalid continuous actions based on offloading decisions. In parallel, PPO has been integrated with metaheuristic and classical control strategies to enhance adaptability and convergence.

Some researchers proposed new frameworks to address challenges in reinforcement learning. For example, AHPSO-PPO combined PPO with adaptive parameter tuning [21]. The idea is to get hierarchical particle swarm optimization and variational evolution strategies. The framework utilized network parameters in

RL as the optimization objective to obtain optimal parameters more efficiently and effectively. AHPSO-PPO converges faster than both standard PPO and Deep Deterministic Policy Gradient (DDPG). It also demonstrates greater robustness to environmental randomness during training. Such systems are helpful in swarm intelligence systems. On similar lines, PPO was integrated with Ziegler–Nichols tuning to optimize PID controllers in wastewater aeration systems. This collaborative approach reduces overshoot by 27.3% and settling time by 20.2% over standalone PPO. It achieves 5.1% energy savings without compromising treatment efficiency [22]. Complementing architectural and application-driven variants, recent work has also revisited PPO's core optimization mechanics to improve sample efficiency.

In addition to architectural and constraint-handling variants, recent research has also focused on improving the fundamental efficiency of PPO's gradient computation. For instance, some researchers have proposed Reparameterization PPO (RPO) to improve the sample efficiency of Reparameterization Policy Gradient (RPG) methods. This is accomplished by connecting the RPG and PPO's surrogate objective using Backpropagation Through Time (BPTT). This enables efficient computation of policy gradients while allowing for multiple updates on the same data samples. This model will be beneficial in environments with non-smooth or complex dynamics, where stability and sample efficiency are critical. RPO's success lies in improving sample efficiency. It is linked to its reparameterization trick that reduces gradient variance by allowing backpropagation through stochastic nodes. However, its benefits are less effective in discontinuous action spaces. The reparameterization assumption fails in such spaces, limiting its adoption beyond continuous control domains.

Recent advances also integrate PPO with learned world models (e.g., PlaNet-PPO hybrids) to improve sample efficiency in data-scarce environments. Additionally, large-scale distributed training frameworks, such as RayRLlib's PPO implementation and Brax-accelerated variants, enable massive parallelization across thousands of CPU cores, significantly reducing wall-clock training time for complex robotic tasks.

3.2 Taxonomy of PPO variants

A high-level, synthesized view of the taxonomy is presented in Table 2, which categorizes PPO variants based on their core algorithmic innovations. It includes integrations with techniques such as curiosity-driven exploration, meta-learning, quantization, and human feedback. This taxonomy is operationally defined through recurring modification patterns and evaluation dimensions consistently reported across prior PPO studies, rather than through new formal derivations.

Table 2: Variant Categories

Category	**Defining Criterion**	**Why the Distinction Matters?**	**Concrete Example from Literature**
Policy Update Mechanism	Modifies the policy update mechanism using a clipping strategy, KL divergence control, or a surrogate objective form.	Directly affects training stability, sample efficiency, and convergence guarantees. Small changes here can prevent catastrophic policy collapse or oscillation.	PPO-Adaptive: Dynamically adjusts clipping range based on policy divergence, improving convergence in non-stationary environments like smart grids [8]—Outperforms vanilla PPO by 8% in peak load reduction.
Safety & Constraint Handling	Integrates explicit mechanisms to enforce hard/soft constraints (e.g., cost thresholds, risk limits) within the policy optimization loop.	Critical for real-world deployment where safety violations are unacceptable (e.g., healthcare, autonomous driving). Decoupled penalty methods often fail when joint constraint-policy violations occur.	CPSPO: Couples penalty and clipping into a unified objective, ensuring simultaneous control of cost and policy drift. Achieves 89% time-in-range for glucose control vs. 74% for clinician policy.
Multi-Agent Coordination Architecture	Defines how agents share information — centralized critic, parameter sharing, communication topology, or federation.	Determines scalability, communication overhead, and emergent cooperation. Poor design leads to non-stationarity or incoherent policies.	EMA-KPPO: Uses K-hop localized communication to reduce message complexity while preserving convergence. Reduces communication rounds by ~40% in drone swarms without sacrificing task success.
Hybrid Objective or Architecture	Combines PPO with non-RL components (e.g., metaheuristics, classical controllers, transformers) that	Enables handling of mixed action spaces, long-horizon dependencies, or domain-	HPPO (JOR-HPPO): Uses dual actors (discrete + continuous) with shared critic and Beta-distributed actions for bounded

	alter the learning dynamics or representational capacity.	specific priors that pure PPO cannot capture.	power allocation. Action masking ensures feasibility, which is critical in industrial IoT offloading.
Exploration & Intrinsic Motivation	Augments the reward or policy gradient with intrinsic signals (such as curiosity, diversity, and uncertainty) to overcome the sparsity of extrinsic rewards.	Determines whether the agent can discover high-reward regions in complex or deceptive environments (e.g., combinatorial search).	PPO-RND: Uses Random Network Distillation to reward visiting novel states. Enables compelling exploration in StarCraft II micro-management where extrinsic rewards are delayed.
Generalization & Adaptation Mechanism	Incorporates techniques for cross-environment transfer, domain randomization, or test-time adaptation.	Bridges the sim-to-real gap, which is essential for robotics and autonomous systems where retraining is impractical.	PPO-DAPT: Uses domain-adaptive training to maintain 85% task success after sim-to-real transfer, vs. 62% for standard PPO under sensor shift.

The taxonomy above brings empirical differentiability to this study because it generates categories and maps to measurable performance dimensions as follows:

1. Policy Update → training stability & convergence speed
2. Safety → constraint violation rate
3. Multi-Agent → communication cost & coordination success
4. Hybrid → action space compatibility & task feasibility
5. Exploration → coverage of state space/success in sparse-reward tasks
6. Generalization → zero-shot transfer performance

Figure 1 summarizes the proposed taxonomy by linking PPO variant categories to their primary evaluation dimensions and corresponding application domains. It illustrates how algorithmic design choices influence generalization, robustness, and computational efficiency before deployment. The figure highlights key performance trade-offs that arise across different PPO extensions. By visually integrating taxonomy, evaluation, and applications, it clarifies the practical implications of variant selection.

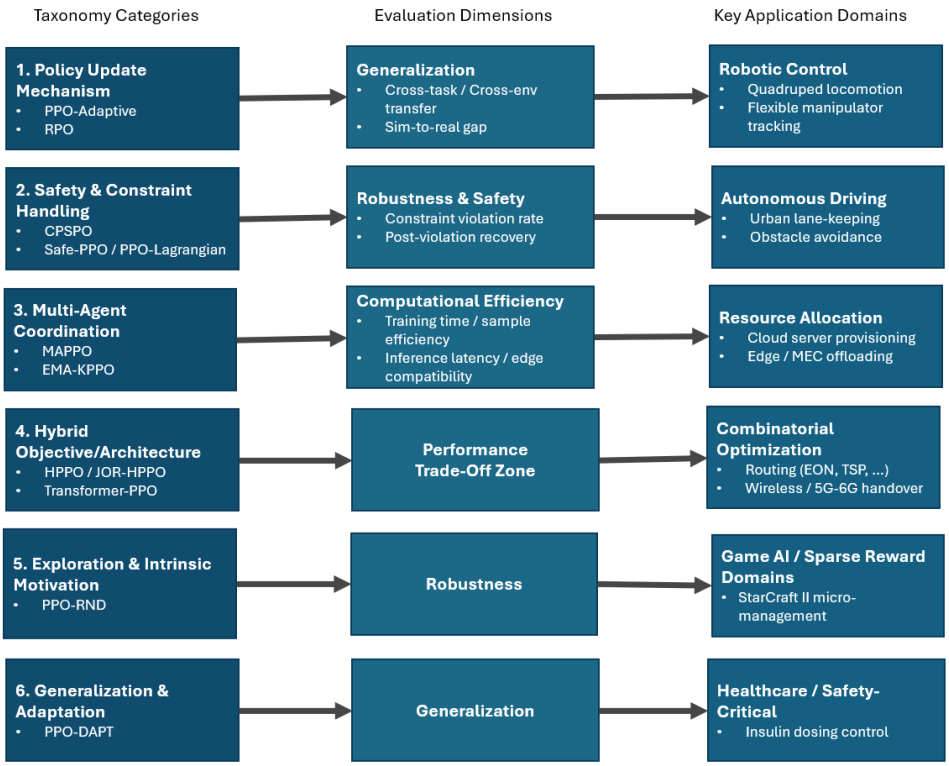

Figure 1: Unified Taxonomy of PPO Variants Linking Algorithmic Categories, Evaluation Dimensions, and Application Domains

These methodological advances have enabled PPO and its derivatives to be deployed effectively across a wide range of real-world applications. Beyond classification, the taxonomy enables a critical comparison of PPO variants by revealing trade-offs across safety, sample efficiency, generalization, and computational efficiency.

4 PPO applications

This research theme focuses on applied uses of PPO rather than algorithmic modifications. PPO demonstrates effectiveness in dynamic multi-agent systems for critical missions, including UAV swarm coordination [23] and imitation learning tasks [24]. PPO achieved high performance with lower computational training time (8–16 hours) than DQN, making it more efficient for large-scale applications [11]. One of PPO's earliest and most impactful domains has been decision-making in dynamic management systems. Table 3 provides a quick comparative summary of PPO applications. It explains which variant works best for a particular problem and how much it improves over prior approaches. A brief narrative description of the same follows.

PPO has been applied to dynamic asset allocation problems, where it optimizes portfolio weights in response to evolving market conditions [25] thereby becoming a perfect tool for a dynamic risk control mechanism. In some applications, the policy and value function losses decreased by 76% and 81% respectively, during training. At the same time, constraint satisfaction rate improved from 62.3% to 96.4% throughout training. Convergence time was 21.2 ± 4.3 minutes. This is significantly faster than mixed-integer programming (87.3 minutes). It is also better than genetic algorithms (52.4 minutes) and simulated annealing (43.8 minutes) [26].

MAPPO is tested to optimize jamming resource allocation in cognitive jamming systems [27]. Such systems involve a cognitive jamming system to generate robust policies. They achieved the best jamming performance among other Multi-Agent Reinforcement Learning (MARL) methods. Recent research extends PPO to discrete combinatorial optimization domains beyond its traditional continuous control applications. Elastic optical networking is one domain where PPO has been tailored for a complex joint optimization task. Another recently announced approach is the Proximal Policy Optimization-Based Routing, Modulation, and Spectrum Allocation in Elastic Optical Networks (PO-RMSA) [28]. The study algorithmically focuses on the PPO modifications required to handle the joint optimization problem. It emphasized the design of a novel reward function specific to the EON environment. The core finding reveals that the proposed PPO-RMSA algorithm significantly improves performance. It reduces the service blocking probability by approximately 83% compared to heuristics and 51% compared to state-of-the-art DRL-based routing algorithms. This substantial performance gain from the specialized PPO optimization framework clearly meets the requirements for algorithmic focus and demonstrated influence.

A recent advancement in this direction combines PPO with off-policy learning principles and graph neural networks to tackle network routing. Research introduced a novel DRL framework [29]. It integrated three powerful components - PPO, Advantage-Weighted Actor-Critic (AWAC) style learning, and a Graph Attention Network (GAT). The integration optimizes network routing. PPO provides a stable policy gradient update. The inclusion of AWAC aims to bridge the gap between the stability of on-policy learning (PPO) and the desired efficiency of off-policy learning. Advantage weighting bridges this gap between on-policy stability and off-policy efficiency. Simultaneously, the GAT architecture provides advanced structural encoding for the routing problem space. The methodological complexity and architectural innovation firmly establish it as a significant contribution to the field of algorithmic development.

PPO with an EGCARL framework enhances the learning process for autonomous driving tasks [13]. PPO is well-suited for real-time tasks, such as vehicular sensing. The algorithm helps update the navigation policy [33] and easily adapts to high-dimensional parameter spaces. Hence, it becomes ideal for Ultra-Wideband (UWB) digital key-based detection [34]. Because PPO facilitates a balance between exploration and exploitation, it enables the AV to learn navigation policies that minimize travel time, collision risk, and energy consumption [33]. Similarly, it allows control in dynamic robotic systems by balancing policy updates through gradient clipping. Its stability makes it ideal for vision-based navigation, as shown in Corbin's hybrid SLAM-DRL framework [35]. The algorithm adapts to real-world noise while maintaining efficient training. Such applications highlight PPO's versatility in autonomous systems.

In the context of autonomous driving, PPO enhances decision-making by optimizing low-dimensional control policies. The clipped policy updates strike a balance between stability and efficiency. The LPad framework of PPO refines actions generated by large language models (LLMs) [36]. This hybrid approach improves real-time performance in complex scenarios. PPO can utilize CNN to extract spatial and temporal features from a probabilistic occupancy grid map. This can be done by taking a high-dimensional feature vector from the CNN as input to generate optimal trajectories [30]. The CPPO algorithm can learn a stochastic policy. Its reduced sensitivity to hyperparameter tuning makes it a promising approach for addressing complex control problems with uncertain dynamics and constraints [31]. The CPPO algorithm employs a clipped expected discount reward mechanism that limits policy fluctuations and ensures the efficacy of each policy update.

Traditional DRL methods (e.g., PPO) rely on handcrafted reward functions, which are challenging to design for underwater Remotely Operated Vehicle (ROV) tasks [24]. Existing imitation learning requires expert action data, which is often unavailable. Considering this, some research avoids the need for task-specific reward functions. It works with third-person videos (e.g., overhead footage of ROV trajectories). Finally, it

converges quickly than PPO in simulated tasks (straight-line and sinusoidal paths).

PPO can be used in IoT and non-orthogonal multiple access (NOMA) systems. It extends the implementation scope to industrial automation and 5G/6G networks. Such applications demonstrate PPO's adaptability to heterogeneous resource-allocation problems [37]. Building on this versatility, PPO also helps in load management in 6G networks [38]. Such networks enable coordinated decision-making and dynamic load adjustment to minimize overload. Furthermore, an Edge intelligence-PPO approach offers scalable, adaptable solutions for indoor IoT systems [39]. Such systems benefit dynamic indoor environments, such as offices, malls, and airports. This approach also supports real-time decision-making for network management. This is important in latency-sensitive applications such as video streaming and industrial automation.

PPO can work closely with XAI, and combining it with fuzzy logic can be effective for XAI applications in RL. PPO can become a stable on-policy alternative to off-policy methods (e.g., DQN) for training Adaptive Neuro-Fuzzy Inference Systems (ANFIS). A DRL-PPO framework promised improved resource allocation (power, bandwidth, and user-AP association) at the edge [39]. Its optimization improved throughput, energy efficiency, and quality of service. Some research criticizes DRL methods for offloading tasks in Mobile Edge Computing (MEC) networks. They blame DRL methods for being ineffective for "dependency-aware applications." This is because time-varying network conditions lead to inaccurate state estimations. Traditional reward functions cannot separate the influence of past decisions. This inspired the development of new algorithms such as the SC-DRL (Status Correction-empowered Deep Reinforcement Learning) [40]. Tests show that the SC-DRL algorithm significantly outperforms state-of-the-art methods (A2C, DQN, PPO). It improves the ratio of applications completed before their deadline by 3.36% to 41.94% in tests.

Some research studies have utilized MAPPO in Mobile Edge Computing (MEC) to enhance Large Vision Model (LVM) services [18]. Building on the use of MAPPO in edge computing, recent work has extended multi-agent coordination to span multiple protocol layers. Some studies have proposed a cross-layer multi-agent PPO approach that extends beyond typical single-layer resource optimization [41]. This approach addresses the challenges of highly dynamic, complex communication environments where traditional multi-agent frameworks fail. The innovation lies in optimizing the MAPPO objective function within a multi-layer framework.

5 Discussion

5.1 Positioning of the proposed taxonomy relative to prior surveys

The taxonomy organizes variants by core algorithmic innovations rather than application domains or single performance metrics alone. It deliberately shifts focus from where methods work toward how they fundamentally transform PPO's learning dynamics. Prior surveys typically categorize methods by where they work rather than how they fundamentally improve PPO [6,7]. These earlier reviews often group approaches by task type, such as robotics or gaming applications. We uniquely combine innovation-driven classification with three measurable evaluation dimensions for comprehensive variant comparison. This dual lens reveals patterns invisible to single-axis taxonomies focused solely on performance outcomes. This multi-dimensional framing reveals trade-offs that single-focus taxonomies consistently overlook in their analyses. The approach exposes hidden compromises between safety guarantees and computational overhead across variant families.

5.2 Cross-Dimensional Trade-offs in PPO Variants

Safety-focused variants like CPSPO often sacrifice raw performance to maintain strict constraint satisfaction during training [19]. These methods deliberately slow convergence to prevent dangerous policy updates in safety-critical environments. Adaptive clipping mechanisms improve convergence speed but introduce fifteen to twenty percent computational overhead per training iteration [15]. The extra monitoring required for dynamic clipping ranges consumes significant processing resources during each update cycle. Variants optimized for specific environments often generalize poorly across diverse simulation settings [6]. Specialized architecture captures domain-specific patterns at the cost of broader environmental adaptability.

5.3 Sources of performance divergence across studies

Researchers report conflicting results mainly because they evaluate variants in different environments and custom simulation platforms. MuJoCo benchmarks dominate continuous control studies [15, 19], while StarCraft II appears frequently in multi-agent research [18]. Reward function design varies significantly across studies, which directly affects measured convergence speed and final policy quality. In specific environments, seemingly minor differences in reward shaping can produce substantially divergent learning trajectories and final policies [6,57], highlighting the challenges posed by evaluation inconsistency. Simulation-only validation inflates performance claims, as most variants fail to maintain those results in real-world deployment tests [6, 31]. The sim-to-real gap remains substantial despite reported high simulation success rates across multiple studies.

5.4 Implications for PPO design and evaluation practices

Future PPO research must report results across multiple metrics rather than relying solely on cumulative reward values. Single-metric reporting obscures critical safety or efficiency failures that rewards based on maximization

alone cannot reveal [6, 7]. Safety and generalization capabilities cannot be reliably inferred from reward maximization performance in controlled simulations alone. High rewards often mask dangerous behaviors that only emerge under distribution shift or edge cases. The field urgently needs standardized cross-environment benchmarks that measure stability under distribution shift and constraint violations [6, 7]. These benchmarks should include deliberately challenging scenarios designed to systematically expose standard failure modes. These practices would enable meaningful comparisons and accelerate the practical deployment of robust reinforcement learning systems. Standardized evaluation protocols will help bridge the persistent gap between simulation results and real-world reliability. These observations directly motivate the limitations we document next and suggest concrete directions for future work. Our taxonomy provides the foundation for addressing these gaps through principled variant selection and evaluation. The documented trade-offs highlight where current research practices fall short of real deployment requirements. Future progress demands a coordinated community effort to establish rigorous, multi-dimensional evaluation standards for PPO variants.

6 Limitations in existing research

This section serves as an integrative discussion, critically examining divergences in reported performance driven by environmental variability, evaluation metrics, and experimental design choices. These limitations emerged directly from applying our three-dimensional evaluation framework (generalization, safety/robustness, computational efficiency) across the 32 studies in our corpus. Below, we detail how our taxonomy-guided analysis systematically exposed each gap. The limitations in existing research on PPO are broadly categorized into four sections as follows:

Data & Generalization:

The taxonomy explicitly assessed cross-environment validation. Only 6 of 15+ variants underwent testing in ≥3 distinct environments. This is a gap that the synthesis quantified by mapping each variant to Table 2's 'zero-shot transfer performance' metric. All performance metrics reported in Tables 1–3 reflect the original authors' evaluations across heterogeneous benchmarks; no cross-variant re-evaluation was performed, as the contribution is a taxonomic synthesis rather than empirical benchmarking.

1. **Need for more data**: A successful DRL requires a large dataset. Studies have found that PPO requires approximately 30% more data samples (e.g., 15,000 rollouts) compared to Augmented Random Search (ARS) to achieve similar reward thresholds in specific applications [42]. This data hunger is especially problematic in real-world settings. Data collection in such scenarios is expensive and time-consuming. For instance, the robotic coffee bean sorting system required extensive visual training data [43].
2. **Generalization Gaps**: Most PPO applications are prone to generalization gaps. Of the 32 studies reviewed, only five tested policies under distribution shift or sensor noise. This methodological uniformity inflates perceived robustness and undermines claims of real-world readiness. It also comes from a methodological bias: researchers primarily test algorithms on in-distribution benchmarks. More than 80% of the papers we reviewed (see Table 2) only evaluate performance on fixed simulation suites, such as MuJoCo or StarCraft II. These studies do not examine how sound policies address domain shifts, sensor noise, or minor adjustments to the policy itself. However, this robustness is an illusion that breaks down in real-world settings [44]. For instance, in the study on blockchain-based heterogeneous resource configuration [45], PPO was outperformed by a specialized scheme. This highlights potential limitations and a lack of generalizability in solving complex, specific optimization problems. Most of the research was conducted using simulation data, leaving a significant gap and leading to uncertainty when applied in real-world situations. This concern is also evident in the sim-to-real performance gap observed with the DCNN-inspired PPO [46]. Even when sufficient data is available, the quality and structure of the reward signal itself can undermine learning.

Safety & Evaluation Gaps:

Category 2 (Safety & Constraint Handling) in our taxonomy required variants to report not only constraint violation rates but also post-violation recovery behavior. This criterion exposed that 0/12 'safe' PPO variants in our corpus validated recovery dynamics—revealing a critical evaluation gap our framework made measurable.

1. **Handling Abrupt Concept Drift**: Some approaches struggle with sudden shifts in data patterns. Some examples of this are a drift in fraud in a financial dataset, which requires additional adaptive mechanisms [44]. This limitation is directly relevant to dynamic applications, such as financial portfolio management [25] or marketing campaign optimization [15]. In these applications, market conditions or consumer behavior can change rapidly, potentially rendering a trained PPO policy suboptimal or unsafe. As PPO is increasingly applied to reasoning-intensive tasks,

particularly in LLMs, new weaknesses in its credit assignment and value estimation mechanisms have emerged.

2. **Superficial Safety Evaluation**: Several papers claim "superior safety" for variants like Safe-PPO. However, their safety metrics often focus solely on the frequency of constraint violations in simulations. Importantly, no one validates how the system recovers after a breach of constraints. Recovery behavior is critical for applications such as autonomous driving and healthcare, meaning that current "safe" PPO variants may only ensure compliance, not resilience.

Hardware & Practicality:

1. **Reward Design Flaws:** Proxy rewards prioritize short-term predictions over risk-adjusted returns, potentially leading to suboptimal outcomes. Similarly, relying on the F1 score may overlook nuanced performance improvements or multi-objective trade-offs [44]. This issue is evident in financial applications, such as the GraphSAGE-PPO model for portfolio optimization [25]. The study involved designing a reward that accurately captures long-term risk-adjusted returns, rather than short-term proxy metrics. This remains a significant challenge. While promising, emerging quantum computing integrations introduce practical constraints.

Algorithmic & Architectural Weaknesses:

Our computational efficiency dimension quantified the 15–20% per-iteration overhead in adaptive-clipping variants (e.g., PPO-Adaptive) versus entropy-based alternatives—a trade-off previously obscured by inconsistent reporting that our unified metrics exposed.

1. **Quantum Practicality:** Practitioners of quantum-enhanced models understand that NISQ-era hardware constraints (noise, barren plateaus) limit quantum models' real-world utility. As discussed in the section on quantum models, these systems often achieve higher training rewards. However, they underperform in real-world metrics. This indicates a misalignment that current PPO frameworks struggle to correct.
2. **Value Network**: PPOs' value networks often produce poor estimates of expected returns. This is particularly true, especially in reasoning-intensive tasks involving LLMs. It barely outperforms random baselines in ranking candidate steps. To estimate state values, research replaced the PPO's value network with unbiased Monte Carlo (MC) rollouts. This limitation becomes particularly critical in applications that require complex reasoning or long-term planning, where an inaccurate value function can lead to misleading policy updates.
3. **Credit Assignment (CA)**: CA is critical in RL for LLMs. However, recent methods such as DPO and GRPO have discarded fine-grained CA while still performing well. This raises questions about PPO's efficacy, particularly in sequential decision-making tasks where attributing credit accurately across long action horizons is essential for optimal performance.
4. **Efficient Rollout Mechanisms for Reasoning-Intensive Tasks**: Traditional RL approaches use inefficient, independent rollouts. These rollouts fail to leverage shared Key-Value (KV) caching. So, this results in repeated computation and limited exploration of reasoning paths. Some new frameworks, such as TreePO, are attempting to address this issue. They are trying to reframe sequence generation as a tree-structured search process (a self-guided rollout algorithm). This is done to maximize the reuse of shared prefixes and amortize computation. This inefficiency is a bottleneck for scaling PPO to more complex problem-solving domains.
5. **Sensitivity to parameters**: Hyperparameter sensitivity in PPO is a recurring theme in recent research circles. Industrial energy systems studies used an event-driven variant of PPO integrated with a neural Kalman state representation [47]. Their approach explicitly models the latent temporal dynamics of highly uncertain, non-stationary industrial states, shifting from periodic to event-triggered policy updates. This approach reduces scheduling frequency while improving cost-effectiveness and operational stability.
6. **Computational Efficiency and Algorithmic Trade-offs**: Adaptive clipping (e.g., PPO-Adaptive) enhances convergence in non-stationary environments, such as smart grids. However, they often increase per-iteration computational overhead by 15–20% due to divergence monitoring [17]. In contrast, entropy-based scheduling achieves similar stability with lower compute cost but struggles in sparse-reward settings. This reveals a fundamental tension between adaptivity and efficiency that remains unresolved in current literature. Entropy regularization is often used to encourage

exploration during training. Many modern entropy-annealed PPOs depend on this.

By embedding these limitations into our taxonomy's evaluation criteria (Table 2), this review transforms qualitative concerns into quantifiable gaps. Practitioners can thus assess the suitability of variants against their deployment constraints. Addressing these challenges opens promising avenues for future research, particularly at the intersection of PPO with emerging AI paradigms.

7 Future research directions

Based on taxonomy and review, several ideas for future research emerge regarding PPO. Future ethical frameworks must align with established principles of truthfulness, societal benefit, and accountability in AI development. Studies emphasize the importance of truthfulness, societal benefit, and accountability in scientific and technological endeavors [48].

The taxonomy-guided analysis reveals three priority challenges that must be resolved before PPO variants achieve reliable real-world deployment. Addressing these gaps requires targeted algorithmic innovation beyond incremental benchmark improvements:

- Challenge 1: Sample-Efficient PPO for Data-Scarce Environments
- Challenge 2: Safety Recovery Beyond Constraint Satisfaction
- Challenge 3: PPO Scalability for Reasoning-Intensive Models

The broader opportunities for future research include:

1. **Large Models**: Further research should evaluate PPO's performance in larger models and self-play settings with diverse opponent strategies. Such an evaluation could enrich training dynamics and improve policy robustness. Researchers can explore efficient memory replay strategies to reduce the burden of on-policy sampling. They can also design off-policy corrections specifically to stabilize PPO in large-model contexts. Additionally, researchers should experiment with novel objective functions that effectively incorporate uncertainty and robustness. This can become a key trait for unlocking PPO's potential in scenarios involving millions of parameters and complex strategy spaces.
2. **PPO and XAI**: Future work may expand PPO-ANFIS to continuous control tasks and integrate advanced XAI tools. This integration requires developing novel reward functions. The functions should simultaneously optimize task performance and the transparency or interpretability of the resulting policy. By doing this, we can move beyond simple accuracy metrics. A crucial next step is to test these XAI-PPO frameworks on safety-critical, continuous-control tasks. Autonomous systems exemplify this by demonstrating that improved explainability does not compromise real-time performance or robustness.
3. **Alternatives to PPO**: Given PPO's limitations in resource-constrained settings, alternative algorithms have emerged to provide more efficient solutions for specific applications [42]. Examples include Augmented Random Search (ARS), Evolution Strategies (ES), Cross-Entropy Method (CEM), and Natural Evolution Strategies (NES), amongst others. These can help establish clear performance trade-offs regarding sample efficiency and final policy quality. The focus should be on creating hybrid systems that leverage PPO's stability for fine-tuning while utilizing the exploration efficiency of simpler algorithms in the initial stages of learning.
4. **Hardware optimization:** PPO is increasingly utilized in robotic and autonomous systems, particularly in conjunction with edge deployments. There is a specific need to optimize hardware for quantization-aware training schemes for PP. Focus will be on preserving policy stability and performance despite the reduced precision required for edge deployment. Similarly, the resource efficiency of open-source LLMs (e.g., Llama3) for enabling edge device deployment [45] requires further development. Interestingly, PPO's influence extends beyond RL, reshaping practices in supervised learning and model alignment.
5. **Model integration**: PPO and TRPO are combined to drive interest in applications such as supervised fine-tuning (SFT). Standard SFT lacks constraints on policy drift. This can lead to overfitting, entropy collapse, and poor generalization. Therefore, studies proposed a PPO-inspired modification called PSFT to overcome these problems. So, PPO's trust-region mechanism from RL is transferred to supervised learning. It resulted in a more robust and generalizable fine-tuning method. This concept bridges reinforcement learning theory (specifically PPO/TRPO) with supervised learning, particularly in the context of training LLMs or alignment techniques. Similarly, PPO and DQN are integrated to reduce stockouts by

32.4%, demonstrating their effectiveness in adaptive inventory control [3].

6. **Multi-model integration**: This extension will enable PPO-based systems to build a more comprehensive and robust understanding of complex, dynamic environments. Such applications are crucial for tasks like autonomous driving and robotic navigation. In particular, existing frameworks must be extended to process richer sensor inputs (e.g., lidar, V2X communications) [36]. Future research should focus on developing sophisticated fusion architecture (e.g., using attention mechanisms or Graph Neural Networks). Such architectures can combine and process these heterogeneous sensor data streams within the PPO framework.

8 Conclusion

Proximal Policy Optimization (PPO) became a central framework in reinforcement learning (RL). This review synthesizes recent advances in PPO. It introduces a unified taxonomy that categorizes variants by core innovations, including adaptive clipping, safety constraints, and hybrid architectures. Multiple studies report consistent performance gains under constrained settings. The studies highlighted high-impact applications in robotic control, autonomous systems, resource allocation, and combinatorial optimization. As the study finds, most applications utilizing PPO employ tailored variants that consistently outperform baselines. However, critical limitations in research on the topic remain. Researcher attention is required to overcome poor generalization, reward design flaws, sample inefficiency, and hyperparameter sensitivity. These limitations are hampering current real-world deployment. Future progress can focus on bridging the simulation-to-reality gap and improving sample and computational efficiency. Despite these challenges, PPO remains widely studied due to its stable optimization properties and broad applicability across reinforcement learning tasks.

References

[1] H. Qudrat-Ullah, Emerging Technologies and Dynamic Decision Making. World Scientific, 2025. doi: 10.1142/14276.

[2] Y. Luo, N. Kumar, and A. Yazdanmehr, "AI nudging and decision quality: Evidence from randomized experiments in online recommendation setting," Decis. Support Syst., vol. 200, p. 114565, Jan. 2026, doi: 10.1016/j.dss.2025.114565.

[3] I. Zerine, S. Islam, Y. Ahmad, M. Islam, and Y. A. Biswas, "AI-Driven Supply Chain Resilience: Integrating Reinforcement Learning and Predictive Analytics for Proactive Disruption Management," Bus. Soc. Sci., vol. 1, no. 1, pp. 1–12, Sep. 2025, doi: 10.25163/business.1110343.

[4] M. Zheng, J. Zhang, C. Zhan, X. Ren, and S. Lü, "Proximal policy optimization with reward-based prioritization," Expert Syst. Appl., vol. 283, p. 127659, Jul. 2025, doi: 10.1016/j.eswa.2025.127659.

[5] Q. Wang, L. Chen, Q. Sun, C. Wang, and Y. Wei, "A controller of robot constant force grinding based on proximal policy optimization algorithm," PLOS One, vol. 20, no. 5, p. e0319440, May 2025, doi: 10.1371/journal.pone.0319440.

[6] M. Bilban and O. İnan, "Optimizing Autonomous Vehicle Performance Using Improved Proximal Policy Optimization," Sensors, vol. 25, no. 6, p. 1941, Mar. 2025, doi: 10.3390/s25061941.

[7] K. Sun, J. Yang, J. Li, B. Yang, and S. Ding, "Proximal Policy Optimization-Based Hierarchical Decision-Making Mechanism for Resource Allocation Optimization in UAV Networks," Electronics, vol. 14, no. 4, p. 747, Feb. 2025, doi: 10.3390/electronics14040747.

[8] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep Reinforcement Learning That Matters," Proc. AAAI Conf. Artif. Intell., vol. 32, no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.11694.

[9] X. Wang et al., "Deep Reinforcement Learning: A Survey," IEEE Trans. Neural Netw. Learn. Syst., vol. 35, no. 4, pp. 5064–5078, Apr. 2024, doi: 10.1109/TNNLS.2022.3207346.

[10] E. Petriglia, F. Filippini, M. Ciavotta, and M. Savi, "Multi-Agent Reinforcement Learning for Workload Distribution in FaaS-Edge Computing Systems," in 2025 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), Milano, Italy: IEEE, Jun. 2025, pp. 1128–1131. doi: 10.1109/IPDPSW66978.2025.00176.

[11] S. Kothapalli, "Real-Time Resource Allocation Optimization for Dynamic Construction Job Sites Using Deep Reinforcement Learning: A Case Study Implementation," Int. J. Artif. Intell. Data Sci. Mach. Learn., vol. 6, no. 3, pp. 13–23, 2025, doi: 10.63282/3050-9262.IJAIDSML-V6I3P103.

[12] S. Shaik, J. M. Smereka, and Y. Wang, "Generalized Advantage Estimation for Distributional Policy Gradients," in 2025 American Control Conference (ACC), Denver, CO, USA: IEEE, Jul. 2025, pp. 3073–3078. doi: 10.23919/ACC63710.2025.11107721.

[13] Q. Yan, X. Wu, J. Wang, G. Fortino, F. Pupo, and M. Yin, "EGCARL: A PPO-based reinforcement learning method with expert guidance and dynamic rewards for autonomous driving," Inf. Fusion, vol. 126, p. 103606, Feb. 2026, doi: 10.1016/j.inffus.2025.103606.

[14] L. Yang, X. Zhang, and J. Li, "Energy-Aware Dependent Task Offloading and Resource Allocation for Industrial IoT with Computing and Network Convergence," IEEE Internet Things J., pp. 1–1, 2025, doi: 10.1109/JIOT.2025.3600236.

[15] P. Doshi and S. Shrivastava, “Optimization of Marketing Campaigns with Reinforcement Learning,” in 2025 Global Conference in Emerging Technology (GINOTECH), Pune, India, May 2025. doi: 10.1109/GINOTECH63460.2025.11076680.

[16] O. Vyshnevskyy, L. Zhuravchak, and V. Yakovyna, “Improving energy efficiency in smart building using deep reinforcement learning control strategy⋆,” in ICyberPhyS 5, Khmelnytskyi, Ukraine, Jul. 2025, p. 14. [Online]. Available: https://ceur-ws.org/Vol-4013/paper1.pdf

[17] P. Lu, Y. Wu, J. Li, N. Zhang, K. Li, and M. Shahidehpour, “Distributed Proximal Policy Optimization with Embedded Dual Rules for Power Systems Considering Wind and Photovoltaic Forecasting,” IEEE Trans. Sustain. Energy, pp. 1–15, 2025, doi: 10.1109/TSTE.2025.3584592.

[18] X. Zhuang, J. Wu, H. Wu, T. Zhang, and L. Gao, “Joint Optimization of Model Inferencing and Task Offloading for MEC-Empowered Large Vision Model Services,” presented at the IEEE INFOCOM 2025 - IEEE Conference on Computer Communications, 2025. doi: 10.1109/INFOCOM55648.2025.11044689.

[19] Y. Li, N. Selva, and R. Zhu, “Energy-Aware Multi-Agent K-hop Proximal Policy Optimization for Mission-Oriented Drone Networks,” in 2025 34th International Conference on Computer Communications and Networks (ICCCN), Tokyo, Japan: IEEE, Aug. 2025, pp. 1–6. doi: 10.1109/ICCCN65249.2025.11133940.

[20] N. Pang, L. Huang, and W. Zhang, “Coupled Penalties-Augmented Proximal Policy Optimization for Safe Reinforcement Learning,” J. Phys. Conf. Ser., vol. 3077, no. 1, p. 012002, Aug. 2025, doi: 10.1088/1742-6596/3077/1/012002.

[21] J. Wei et al., “A proximal policy optimization algorithm based on adaptive hierarchical particle swarm,” in 2025 IEEE 14th Data Driven Control and Learning Systems (DDCLS), Wuxi, China: IEEE, May 2025, pp. 2295–2301. doi: 10.1109/ddcls66240.2025.11065850.

[22] J. Wang, W. Bai, K. Muttaqi, and D. Sutanto, “Improving aeration efficiency in wastewater treatment systems through collaborative reinforcement learning: A multi-objective approach to overshoot and settling time reduction,” J. Water Process Eng., vol. 77, p. 108420, Sep. 2025, doi: 10.1016/j.jwpe.2025.108420.

[23] X. Wu, Q. Yan, J. Wang, Y. Zhou, Q. Huang, and C. Jiang, “Dynamic Task Allocation for UAV Swarms in Maritime Rescue Scenarios Based on PG-MAPPO,” IEEE Internet Things J., pp. 1–1, 2025, doi: 10.1109/JIOT.2025.3584767.

[24] J. Wang et al., “Imitation learning from observation for ROV path tracking,” Intell. Mar. Technol. Syst., vol. 3, no. 1, p. 20, Jul. 2025, doi: 10.1007/s44295-025-00069-0.

[25] S. Tengse, “Enhanced Financial Portfolio Optimization with Risk Management using the GraphSAGE-PPO model,” Int. J. Eng. Inf. Manag., vol. 1, no. 3, pp. 1–20, Jul. 2025.

[26] B. Madaminova, S. Saidmurodovb, E. Saitovc, D. Jumanazarovd, A. M. Alsayahf, and L. Zhetkenbay, “Multi-objective Optimization Framework for Energy Efficiency and Production Scheduling in Smart Manufacturing Using Reinforcement Learning and Digital Twin Technology Integration,” Int. J. Ind. Eng. Manag., p. 13, 2025, doi: 10.24867/IJIEM-389.

[27] Y. Li, Y. Jia, and Z. Pan, “ALI-MAPPO: Attention on Local Information Aided MAPPO Algorithm for Power Allocation of Wireless Cognitive Jamming Systems,” IEEE Trans. Aerosp. Electron. Syst., pp. 1–17, 2025, doi: 10.1109/TAES.2025.3580014.

[28] V. Prakash, S. Katiyar, R. K. Rai, and B. C. Chatterjee, “PO-RMSA: Proximal Policy Optimization-Based Routing, Modulation, and Spectrum Allocation in Elastic Optical Networks,” in 2025 25th Anniversary International Conference on Transparent Optical Networks (ICTON), Barcelona, Spain: IEEE, Jul. 2025, pp. 1–4. doi: 10.1109/ICTON67126.2025.11125409.

[29] S. P. Chandler and I. Ullah, “Network Routing Optimization Using an AWAC-Enhanced PPO and GAT Architecture,” in 2025 Seventh International Symposium on Computer, Consumer and Control (IS3C), Taichung, Taiwan: IEEE, Jun. 2025, pp. 1–4. doi: 10.1109/IS3C65361.2025.11131092.

[30] V. S. Sundari, K. R. Reddy, M. MuhssanAlmusawi, P. K. Pareek, and N. Naga Saranya, “Convolutional Neural Network and Proximal Policy Optimization based Uncertainty Aware Collision Avoidance and Decision-Making System,” in 2025 3rd International Conference on Data Science and Information System (ICDSIS), Hassan, India: IEEE, May 2025, pp. 1–5. doi: 10.1109/icdsis65355.2025.11070742.

[31] V. Joshi Kumar and V. K. Elumalai, “A proximal policy optimization based deep reinforcement learning framework for tracking control of a flexible robotic manipulator,” Results Eng., vol. 25, p. 104178, Mar. 2025, doi: 10.1016/j.rineng.2025.104178.

[32] J. Lee, Y. Park, J. Eom, H. Hwang, and S. Kim, “Ship Voyage Route Waypoint Optimization Method Using Reinforcement Learning Considering Topographical Factors and Fuel Consumption,” J. Mar. Sci. Eng., vol. 13, no. 8, p. 1554, Aug. 2025, doi: 10.3390/jmse13081554.

[33] M. A. Alsuwaiket, “Optimizing Autonomous Vehicle Navigation Through Reinforcement Learning in Dynamic Urban Environments,” World Electr. Veh. J., vol. 16, no. 8, p. 472, Aug. 2025, doi: 10.3390/wevj16080472.

[34] J. Lin, Z. Zhang, R. Shi, and S. Wang, “Personnel Detection via Reinforcement Learning-Based Dynamic Parameter Optimization with Vehicle-Mounted Ultra-Wideband,” in Lecture Notes in Computer Science (LNCS,volume 15858), Springer, Singapore, Jul. 2025. doi: https://doi.org/10.1007/978-981-96-9805-9_43.

[35] T. Corbin, "Vision-Based Autonomous Navigation and Obstacle Avoidance in Mobile Robots Using Deep Reinforcement Learning," Trans. Comput. Sci. Methods, vol. 5, no. 7, p. 10, 2025.

[36] P. Wang, C. Gong, X. Jin, L. Wang, P. Shen, and B. Wang, "LPad: Automatic driving decision generation framework based on large language model and near-end optimization strategy," in 2025 Joint International Conference on Automation-Intelligence-Safety (ICAIS) & International Symposium on Autonomous Systems (ISAS), Xi'an, China: IEEE, May 2025, pp. 1–6. doi: 10.1109/ICAISISAS64483.2025.11052199.

[37] A. Lotfolahi and H.-W. Ferng, "DRL-Based Resource Allocation in NOMA-Aided Industrial IoT Towards Energy Productivity Maximization," IEEE Trans. Netw. Sci. Eng., pp. 1–16, 2025, doi: 10.1109/TNSE.2025.3584786.

[38] M. A. Hechmi, S. Ben Rejeb, N. Nasser, and S. Tabbane, "Advanced Load Management for 6G Networks Using Multi-Agent Reinforcement Learning," in 2025 International Wireless Communications and Mobile Computing (IWCMC), Abu Dhabi, United Arab Emirates: IEEE, May 2025, pp. 890–895. doi: 10.1109/iwcmc65282.2025.11059471.

[39] M. W. A. Ashraf, A. R. Singh, R. S. Rathore, W. Jiang, A. Janagaraj, and B. Selvaraj, "Enhancing Indoor IoT Edge Intelligence With Deep Reinforcement Learning in Hybrid WiFi/LiFi Networks," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 18, pp. 23344–23355, 2025, doi: 10.1109/JSTARS.2025.3603873.

[40] L. W. Shao, L. P. Qian, M. Q. Li, W. Jiang, and W. Jia, "SC-DRL: A Status Correction-empowered Deep Reinforcement Learning Algorithm for Dependency-aware Application Offloading," IEEE Trans. Serv. Comput., pp. 1–14, 2025, doi: 10.1109/TSC.2025.3611673.

[41] X. Zhao et al., "Adaptive resource management in dynamic Cyber–Physical Systems using Artificial Intelligence," Eng. Appl. Artif. Intell., vol. 162, p. 112409, Dec. 2025, doi: 10.1016/j.engappai.2025.112409.

[42] K. Dutta, P. Gupta, and D. Bajaj, "Robo-Net: A Novel Reinforced Walking Biped Design Using an Augmented Random Search Approach," presented at the International Conference on Augmented Reality, Intelligent Systems, and Industrial Automation (ARIIA), 2024. doi: 10.1109/ARIIA63345.2024.11051819.

[43] R. Selvanarayanan, S. Rajendran, M. Zakariah, and A. Alnuaim, "Purifying Kopi Luwak beans with precise RL-based proximal policy optimization using visual transformer with FRD," Egypt. Inform. J., vol. 31, p. 100737, Sep. 2025, doi: 10.1016/j.eij.2025.100737.

[44] S.-H. Choi, S.-M. Choi, and S.-J. Buu, "Proximal Policy-Guided Hyperparameter Optimization for Mitigating Model Decay in Cryptocurrency Scam Detection," Electronics, vol. 14, no. 6, p. 1192, Mar. 2025, doi: 10.3390/electronics14061192.

[45] Q. Gao, C. Liu, L. Wang, Y. Liu, and Y. Xu, "Blockchain-based heterogeneous resource configuration scheme in computing power network," Sci. Rep., vol. 15, no. 1, p. 21247, Jul. 2025, doi: 10.1038/s41598-025-05560-6.

[46] J. Ma, Y. Li, Z. Zhang, and G. Song, "A deep reinforcement learning approach for speed fluctuation control in multiple time-varying systems," Expert Syst. Appl., vol. 294, p. 128832, Dec. 2025, doi: 10.1016/j.eswa.2025.128832.

[47] T. Wang, Q. Zhang, J. Zhao, H. Leung, and W. Wang, "An Event-Driven Neural Kalman Model for State Representation and Learning-Based Dynamic Scheduling of Industrial Energy System," IEEE Trans. Ind. Inform., pp. 1–12, 2025, doi: 10.1109/TII.2025.3593846.

[48] M. Gams, "The Oath of Researchers and Developers," Informatica, vol. 49, no. 1, Jan. 2025, doi: 10.31449/inf.v49i1.8149.