Novelty Assessment Report

Paper: Safe Exploration via Policy Priors

PDF URL: https://openreview.net/pdf?id=JC8xYAADHL

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-27

Abstract

Safe exploration is a key requirement for reinforcement learning agents to learn and adapt online, beyond controlled (e.g. simulated) environments. In this work, we tackle this challenge by utilizing suboptimal yet conservative policies (e.g., obtained from offline data or simulators) as priors. Our approach, SOOPER, uses probabilistic dynamics models to optimistically explore, yet pessimistically fall back to the conservative policy prior if needed. We prove that SOOPER guarantees safety throughout learning, and establish convergence to an optimal policy by bounding its cumulative regret. Extensive experiments on key safe RL benchmarks and real-world hardware demonstrate that SOOPER is scalable, outperforms the state-of-the-art and validate our theoretical guarantees in practice.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **safe exploration in reinforcement learning with policy priors**

A total of **35 papers** were analyzed and organized into a taxonomy with **13 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Policy Prior Integration Mechanisms**
- **Constraint-Based Safe Learning**
- **Exploration Strategies with Safety Guarantees**
- **Domain-Specific Safe RL Applications**

Complete Taxonomy Tree

- safe exploration in reinforcement learning with policy priors Survey Taxonomy
- Policy Prior Integration Mechanisms
  - Controller Fusion and Hybrid Architectures (3 papers)
  - [2] Bayesian controller fusion: Leveraging control priors in deep reinforcement learning for robotics (Rana, 2023) View paper
  - [18] Multiplicative Controller Fusion: Leveraging Algorithmic Priors for Sample-efficient Reinforcement Learning and Safe Sim-To-Real Transfer (Krishan Rana, 2020) View paper
  - [31] Constrained Residual Race: An Efficient Hybrid Controller for Autonomous Racing (Mengmeng Zou, 2023) View paper
  - Policy Transfer and Adaptation (4 papers)
  - [6] Safe Adaptive Policy Transfer Reinforcement Learning for Distributed Multiagent Control (Bin Du, 2023) View paper
  - [10] Learning transferable domain priors for safe exploration in reinforcement learning (Karimpanal, 2020) View paper
  - [27] Probabilistic policy reuse for safe reinforcement learning (Javier Garcia, 2019) View paper
  - [30] Training and Transferring Safe Policies in Reinforcement Learning (Q, 2022) View paper
  - Prior-Guided Policy Initialization (4 papers)
  - [11] Jump-Start Reinforcement Learning with Self-Evolving Priors for Extreme Monopedal Locomotion (Zhan, 2025) View paper
  - [14] Rocket Landing Control with Random Annealing Jump Start Reinforcement Learning (Yuxuan Jiang, 2024) View paper
  - [15] Coordination of Bounded Rational Drones Through Informed Prior Policy (Durgakant Pushp, 2023) View paper
  - [17] Planning with RL and episodic-memory behavioral priors (Beohar, 2022) View paper
- Constraint-Based Safe Learning
  - Constrained Optimization and Primal-Dual Methods (4 papers)
  - [1] Real-Time Optimal Power Flow Method via Safe Deep Reinforcement Learning Based on Primal-Dual and Prior Knowledge Guidance (Pengfei Wu, 2025) View paper
  - [21] Accelerating Safe Reinforcement Learning with Constraint-mismatched Policies (Yang, 2022) View paper
  - [22] Anytime-competitive reinforcement learning with policy prior (Yang Jian-yi, 2023) View paper
  - [25] Accelerating Safe Reinforcement Learning with Constraint-mismatched Baseline Policies (Jimmy Jimmy, 2021) View paper
  - Expert Guidance and Demonstration-Based Safety (3 papers)
  - [19] Learning safe policies with expert guidance (Huang Jessie, 2018) View paper
  - [34] Directed Policy Gradient for Safe Reinforcement Learning with Human Advice (Plisnier, 2018) View paper
  - [35] From Game-Playing to Self-Driving: Comparing AlphaGo vs AlphaZero Approaches for Driving Controls (E Xu, n.d.) View paper
  - Adaptive Regularization and Sample Manipulation (2 papers)
  - [7] Enhancing efficiency of safe reinforcement learning via sample manipulation (Yuhao Ding, 2024) View paper
  - [8] Reinforcement learning with adaptive regularization for safe control of critical systems (Pietro Ferraro, 2024) View paper
- Exploration Strategies with Safety Guarantees
  - Optimistic-Pessimistic Exploration with Policy Priors ★ (2 papers)
  - [0] Safe Exploration via Policy Priors (Anon et al., 2026) View paper

  - [26] Safetyâ□□Efficiency Balanced Navigation for Unmanned Tracked Vehicles in Uneven Terrain Using Prior-Based Ensemble Deep Reinforcement Learning (Yiming Xu, 2025) View paper
  - Risk-Aware and Uncertainty-Based Methods (3 papers)
  - [3] Safe exploration for interactive machine learning (Turchetta, 2019) View paper
  - [13] Safe Exploration Method for Reinforcement Learning under Existence of Disturbance (Yoshihiro Okawa, 2022) View paper
  - [28] Amortized Safe Active Learning for Real-Time Data Acquisition: Pretrained Neural Policies from Simulated Nonparametric Functions (Li, 2025) View paper
  - Adaptive Safe Action Selection (2 papers)
  - [9] Trajectory-wise iterative reinforcement learning framework for auto-bidding (Haoming Li, 2024) View paper
  - [29] TempoRL: Temporal Priors for Exploration in Off-Policy Reinforcement Learning (Bagatella, 2021) View paper
  - Reward-Free and Batch Safe Exploration (2 papers)
  - [5] Provably good batch off-policy reinforcement learning without great exploration (Yao Liu, 2020) View paper
  - [33] Safe Exploration Incurs Nearly No Additional Sample Complexity for Reward-free RL (Huang, 2022) View paper
- Domain-Specific Safe RL Applications
  - Robotics and Autonomous Systems (4 papers)
  - [4] Quantum computing and neuromorphic computing for safe, reliable, and explainable multi-agent reinforcement learning: optimal control in autonomous robotics (Taghavi, 2024) View paper
  - [12] Specialized Deep Residual Policy Safe Reinforcement Learning-Based Controller for Complex and Continuous State-Action Spaces (Ammar N. Abbas, 2023) View paper
  - [24] Safe exploration in reinforcement learning: Theory and applications in robotics (Berkenkamp, 2019) View paper
  - [32] Learning to navigate through abstraction and adaptation (Hutsebaut-Buysse, 2023) View paper
  - Autonomous Driving and Vehicle Control (1 papers)
  - [16] Reliable Reinforcement Learning for Decision-Making in Autonomous Driving (Lu, 2024) View paper
  - Generative Models and Latent Variable Methods (2 papers)
  - [20] Training and Evaluation of Deep Policies using Reinforcement Learning and Generative Models (Ghadirzadeh, 2022) View paper
  - [23] Data-efficient visuomotor policy training using reinforcement learning and generative models (Ghadirzadeh, 2020) View paper

Narrative

Core task: safe exploration in reinforcement learning with policy priors. The field addresses how agents can learn effectively while respecting safety constraints by leveraging prior knowledge or baseline policies. The taxonomy reveals four main branches that capture complementary perspectives on this challenge. Policy Prior Integration Mechanisms examines architectural and algorithmic strategies for combining learned policies with existing controllers or expert demonstrations, including methods like controller fusion and residual policy learning. Constraint-Based Safe Learning focuses on formulating and enforcing explicit safety constraints during training, often through optimization frameworks that balance performance objectives with hard or soft safety requirements. Exploration Strategies with Safety Guarantees develops principled approaches to guide exploration toward informative yet safe regions of the state-action space, sometimes employing optimistic-pessimistic trade-offs or probabilistic safety certificates. Domain-Specific Safe RL Applications translates these ideas into concrete settings such as robotics, autonomous driving, and power systems, where real-world consequences demand reliable safety assurances.

Within the exploration strategies branch, a particularly active line of work investigates how to balance optimism for learning with pessimism for safety, often using policy priors as anchors to prevent catastrophic failures. Safe Exploration Policy Priors[0] sits squarely in this optimistic-pessimistic exploration cluster, sharing thematic connections with Safety-Efficiency Balanced Navigation[26], which similarly navigates the tension between task performance and constraint satisfaction. Nearby efforts like Safe Interactive Learning[3] and Efficient Safe RL Sampling[7] emphasize sample-efficient exploration under safety constraints, while works such as Bayesian Controller Fusion[2] and Adaptive Regularization Safe Control[8] blend prior knowledge with adaptive learning mechanisms. The original paper's emphasis on policy priors as a foundation for safe exploration distinguishes it from purely constraint-driven approaches, positioning it as a bridge between integration mechanisms and exploration guarantees that leverages existing knowledge to guide safe discovery of improved behaviors.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Safetyâ□□Efficiency Balanced Navigation for Unmanned Tracked Vehicles in Uneven Terrain Using Prior-Based Ensemble Deep Reinforcement Learning

Authors: Yiming Xu, Dianhao Zhang, Mien Van | **Year/Venue**: 2025 • World Electric Vehicle Journal | **URL**: View paper

Abstract

This paper proposes a novel navigation approach for Unmanned Tracked Vehicles (UTVs) using prior-based ensemble deep reinforcement learning, which fuses the policy of the ensemble Deep Reinforcement Learning (DRL) and Dynamic Window Approach (DWA) to enhance both exploration efficiency and deployment safety in unstructured off-road environments. First, by integrating kinematic analysis, we introduce a novel state and an action space that account for rugged terrain features and trackâ□□ground int...

Relationship Analysis

Both papers belong to the Optimistic-Pessimistic Exploration with Policy Priors category, using conservative priors to maintain safety while exploring. The original paper (SOOPER) focuses on safe exploration in general reinforcement learning with CMDPs using probabilistic dynamics models and policy priors from offline data or simulators, providing theoretical guarantees on cumulative regret. The candidate paper addresses navigation for Unmanned Tracked Vehicles in uneven terrain, combining ensemble Deep RL with Dynamic Window Approach as a behavioral prior, focusing on practical robotic navigation rather than general safe RL theory.

Contributions Analysis

Overall novelty summary. The paper proposes SOOPER, an algorithm that uses probabilistic dynamics models to balance optimistic exploration with pessimistic fallback to conservative policy priors, ensuring safety throughout learning. It resides in the 'Optimistic-Pessimistic Exploration with Policy Priors' leaf, which contains only two papers total (including this one). This indicates a relatively sparse research direction within the broader safe RL landscape, suggesting the specific combination of optimistic-pessimistic balancing with explicit prior fallback mechanisms remains underexplored compared to adjacent areas like constraint-based methods or risk-aware techniques.

The taxonomy reveals that SOOPER's leaf sits within 'Exploration Strategies with Safety Guarantees', a branch containing four leaves and thirteen papers total. Neighboring leaves include 'Risk-Aware and Uncertainty-Based Methods' (three papers focusing on probabilistic safety without explicit prior fallback) and 'Adaptive Safe Action Selection' (two papers emphasizing step-by-step action filtering). The sibling paper in SOOPER's leaf addresses similar optimistic-pessimistic trade-offs but may differ in implementation details or theoretical frameworks. The taxonomy's scope notes clarify that methods without explicit prior fallback mechanisms belong to adjacent categories, positioning SOOPER at a specific intersection of prior-guided learning and exploration guarantees.

Among twenty-four candidates examined, the contribution-level analysis reveals mixed novelty signals. The core SOOPER algorithm (Contribution 1) examined four candidates with zero refutations, suggesting limited direct overlap in algorithmic design. However, the theoretical guarantees (Contribution 2) examined ten candidates and found three potential refutations, indicating that safety and regret bounds in this setting may have substantial prior work. The empirical validation (Contribution 3) examined ten candidates with no refutations, though this likely reflects differences in experimental domains rather than fundamental novelty. The limited search scope means these findings capture top semantic matches, not exhaustive coverage.

Given the sparse taxonomy leaf and limited search scale, SOOPER appears to occupy a relatively underexplored niche combining policy priors with optimistic-pessimistic exploration. The algorithmic contribution shows stronger novelty signals than the theoretical guarantees, where prior work on safety and regret analysis appears more developed. The analysis is constrained by examining only twenty-four candidates from semantic search, leaving open the possibility of additional relevant work outside this scope.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: SOOPER algorithm for safe exploration using policy priors

Description: The authors introduce SOOPER, a model-based reinforcement learning algorithm that uses suboptimal yet conservative policies (obtained from offline data or simulators) as priors. The algorithm employs probabilistic dynamics models to explore optimistically while pessimistically falling back to the conservative policy prior when needed to maintain safety.

This contribution was assessed against **4 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Context-Aware Policy-Guided Gradient Search for Offline Model-Based Optimization

URL: View paper

Brief Assessment

Context-Aware Policy-Guided Search[38] focuses on offline model-based optimization in high-dimensional design spaces, not online safe exploration in reinforcement learning with safety constraints.

2. Reinforcement learning with adaptive regularization for safe control of critical systems

URL: View paper

Brief Assessment

Adaptive Regularization Safe Control[8] focuses on combining MPC-based safety regularizers with model-free RL for critical systems, not on model-based RL with probabilistic dynamics models and optimistic-pessimistic exploration as in SOOPER.

3. Safe model-based reinforcement learning with stability guarantees

URL: View paper

Brief Assessment

Safe Model-Based Stability[36] focuses on Lyapunov-based stability verification for fixed policies with Gaussian process models, not on using policy priors for safe exploration in the manner proposed by SOOPER.

4. A KL-regularization framework for learning to plan with adaptive priors

URL: View paper

Brief Assessment

KL-Regularization Adaptive Priors[37] focuses on model predictive control with learned policy priors for planning in continuous control, not on safe exploration with constraint satisfaction guarantees in CMDPs using pessimistic policy priors from offline data or simulators.

Contribution 2: Theoretical guarantees for safety and cumulative regret bound

Description: The authors provide theoretical analysis proving that SOOPER maintains safety throughout learning with high probability and establishes a novel bound on cumulative regret. This improves over prior works that only guarantee optimality at the end of training, by ensuring good performance during exploration as well.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Regret guarantees for model-based reinforcement learning with long-term average constraints

URL: View paper

Brief Assessment

Long-Term Average Constraints[47] addresses ergodic MDPs with long-term average constraints using posterior sampling, achieving $\tilde{O}(\sqrt{T})$ regret. The original paper focuses on safe exploration in continuous state-action spaces with high-probability safety guarantees throughout learning, using model-based RL with policy priors—a fundamentally different problem setting and approach.

2. Safety and robustness in reinforcement learning

URL: View paper

Brief Assessment

Safety Robustness RL[46] focuses on simple regret rather than cumulative regret. The candidate explicitly distinguishes simple regret from instantaneous regret, whereas the original paper establishes novel bounds on cumulative regret during exploration.

3. Conservative safety critics for exploration

URL: View paper

Brief Assessment

Conservative Safety Critics[43] focuses on bounding the probability of catastrophic failures during training using conservative safety critics, but does not establish cumulative regret bounds during exploration as SOOPER does. The candidate's theoretical guarantees concern failure probability bounds (Theorem 1) and convergence rates (Theorem 2), not cumulative regret during safe exploration.

4. Probabilistic constraint for safety-critical reinforcement learning

URL: View paper

Brief Assessment

Probabilistic Constraint Safety-Critical[44] focuses on probabilistic constraints with high-probability safety guarantees throughout learning, but uses a different formulation (probabilistic constraints vs. SOOPER's pessimistic policy priors). The candidate establishes cumulative regret bounds for their specific probabilistic-constrained setting, which differs from SOOPER's CMDP formulation with policy priors.

5. Safe reinforcement learning in constrained markov decision processes

URL: View paper

Prior Art Analysis

Safe Constrained MDPs[39] demonstrates prior work that provides theoretical guarantees for both safety throughout learning and cumulative regret bounds. The candidate paper explicitly proves that their algorithm guarantees constraint satisfaction with high probability during all episodes and establishes bounds on cumulative regret. This directly challenges the novelty claim that SOOPER is the first to provide such combined guarantees, as the candidate shows similar theoretical results were established earlier.

Evidence

Evidence 1 - **Rationale**: This pair shows that Safe Constrained MDPs[39] already provided theoretical guarantees for both safety and near-optimality (which relates to cumulative regret), challenging the claim that SOOPER is novel in providing such combined guarantees. - **Original**: we prove that sooper guarantees constraint satisfaction with high probability throughout learning. we additionally show that sooper gradually expands an implicit set of safe policies until it converges to the feasible set defined by the true cmdp. this is accomplished by reformulating the problem as... - **Candidate**: we examine sno-mdp by applying pac-mdp analysis and prove that, with high probability, the acquired policy is near-optimal with respect to the cumulative reward while guaranteeing safety.

Evidence 2 - **Rationale**: Both papers provide formal safety guarantees with high probability throughout learning. The candidate's Theorem 1 explicitly states safety is maintained at all timesteps, which is the same type of guarantee claimed as novel in the original paper. - **Original**: theorem 1 (safety guarantee). suppose assumptions 1 to 4 hold and fn is well-calibrated ∀n = 1, . . . , naccording to definition 1. if actions are executed for all timesteps t according to ¯πn(at|st, c<t, qˆπ c,n) := πn(·|st) if φ(st, at, c<t, qˆπ c,n) < d ˆπ(st) otherwise, where φ is the discounted... - **Candidate**: theorem 1. assume that the safety function g satisfies ‖g‖2 k ≤ bg and is l-lipschitz continuous. also, assume that s0̸= ∅ and g(s) ≥ h for all s ∈ s0. fix any ϵg > 0 and ∆g ∈ (0,1). suppose that we conduct the stage of "exploration of safety" with the noise ng t being σg-sub-gaussian, and that βt =...

6. Truly no-regret learning in constrained mdps

URL: View paper

Prior Art Analysis

Truly No-Regret Constrained[45] demonstrates prior work that addresses both safety guarantees throughout learning and cumulative regret bounds. The candidate paper explicitly proves that their algorithm maintains constraint satisfaction with high probability throughout all episodes while establishing sublinear cumulative regret bounds. This directly challenges the novelty claim that SOOPER is the first to provide such combined guarantees, as the candidate paper addresses the same problem of ensuring safety during exploration while bounding cumulative regret.

Evidence

Evidence 1 - **Rationale**: Both papers claim to prove safety throughout learning combined with cumulative regret bounds, indicating the candidate addresses the same theoretical contribution. - **Original**: we prove that sooper guarantees safety throughout learning, and establish convergence to an optimal policy by bounding its cumulative regret - **Candidate**: we prove that our algorithm achieves sublinear regret without error cancellations

Evidence 2 - **Rationale**: The candidate paper explicitly claims to be the first to achieve sublinear regret without error cancellations (which allows safety during learning), directly challenging the original paper's novelty claim of being first to bound cumulative regret while maintaining safety. - **Original**: establish a novel upper bound on the cumulative regret, improving over prior works that provide optimality guarantees only at the end of training with arbitrarily poor performance during exploration - **Candidate**: can we design an efficient primal-dual algorithm that achieves sublinear strong regret in an unknown cmdp? we provide the first affirmative answer for tabular finite-horizon cmdps

Evidence 3 - **Rationale**: Both papers provide formal theorems proving constraint satisfaction with high probability throughout learning combined with regret bounds, showing the candidate addresses the same theoretical guarantee. - **Original**: under standard regularity assumptions, we prove that sooper guarantees constraint satisfaction with high probability throughout learning - **Candidate**: theorem 5.1 (regret bound). let τ = k-1/7, η = (h2i)-1ξk-5/7, λmax = hξ-1k1/14. then with probability at least 1 - δ, algorithm 1 obtains a strong regret of r(k; r) ≤ crk0.93, r(k; u) ≤ cuk0.93

7. Triple-q: A model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation

URL: View paper

Prior Art Analysis

Triple-Q Constrained RL[48] demonstrates prior work that achieves both sublinear cumulative regret and zero constraint violation guarantees throughout learning. The candidate paper explicitly proves that their algorithm maintains safety (zero constraint violation) during all episodes while achieving sublinear cumulative regret bounds. This directly refutes the novelty claim that SOOPER is the first to provide such guarantees, as Triple-Q[48] establishes these properties for constrained reinforcement learning problems.

Evidence

Evidence 1 - **Rationale**: This pair shows that Triple-Q[48] already established sublinear regret with zero constraint violation guarantees, which is the core theoretical contribution claimed by the original paper. - **Original**: we prove that sooper guarantees safety throughout learning, and establish convergence to an optimal policy by bounding its cumulative regret. extensive experiments on key safe rl benchmarks and real-world hardware demonstrate that sooper is scalable, outperforms the state-of-the-art and validate our... - **Candidate**: this paper presents the first model-free, simulator-free reinforcement learning algorithm for constrained markov decision processes (cmdps) with sublinear regret and zero constraint violation. the algorithm is named triple-q because it includes three key components: a q-function (also called actionva...

Evidence 2 - **Rationale**: This evidence pair demonstrates that Triple-Q[48] provides explicit bounds on cumulative regret while maintaining zero constraint violation throughout learning, directly addressing the same theoretical gap that the original paper claims to fill. - **Original**: establish a novel upper bound on the cumulative regret, improving over prior works that provide optimality guarantees only at the end of training with arbitrarily poor performance during exploration (wagener et al., 2021; as et al., 2025b). - **Candidate**: we prove triple-q achieves ˜o (1 δh4s 1 2 a 1 2 k 4 5) reward regret and guarantees zero constraint violation when the total number of episodes k ≥ (16 √ sah6ι3 δ)5 , where ι is logarithmic in k. therefore, in terms of constraint violation, our bound is sharp for large k. to the best of our know...

Evidence 3 - **Rationale**: This pair shows Triple-Q[48] provides formal theoretical guarantees on both cumulative regret and constraint violation, with explicit bounds that ensure safety throughout learning rather than only at convergence. - **Original**: using this formulation, we further under review as a conference paper at iclr 2026 establish a novel upper bound on the cumulative regret, improving over prior works that provide optimality guarantees only at the end of training with arbitrarily poor performance during exploration - **Candidate**: theorem 1. assume k ≥ (16 √ sah6ι3 δ)5 , where ι= 128 log(√ 2sahk). triple-q achieves the following regret and constraint violation bounds: regret(k) ≤13 δ h4√ saι3k0.8 + 4h4ι k1.2 violation(k) ≤54h4ιk0.6 δ log 16h2√ι δ + 4 √ h2ι δ k0.8 -5 √ sah6ι3k0.8. if we further have k ≥e 1 δ, then violation...

8. Adversarially Trained Weighted Actor-Critic for Safe Offline Reinforcement Learning

URL: View paper

Brief Assessment

Adversarially Trained Weighted Actor-Critic[42] focuses on safe offline RL with policy improvement guarantees relative to a reference policy, not online exploration with cumulative regret bounds during learning.

9. Rethinking safe policy learning for complex constraints satisfaction: A glimpse in real-time security constrained economic dispatch integrating energy storage units

URL: View paper

Brief Assessment

Real-Time Security Dispatch[41] focuses on power system economic dispatch with complex constraints, not general RL exploration with cumulative regret bounds during learning.

10. Safe reinforcement learning with contextual information: Theory and applications

URL: View paper

Brief Assessment

Safe RL Contextual Information[40] focuses on contextual information in safe RL with regret analysis, but the provided candidate context is too limited (only fragments) to determine if it addresses the same problem of maintaining safety during exploration with cumulative regret bounds as SOOPER does for policy-prior-based safe exploration.

Contribution 3: Empirical validation on benchmarks and real-world hardware

Description: The authors perform extensive experiments demonstrating that SOOPER outperforms state-of-the-art baselines on standard safe RL benchmarks and validate the approach on real-world robotic hardware, providing empirical evidence that their theoretical guarantees translate to practice.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Robot reinforcement learning on the constraint manifold

URL: View paper

Brief Assessment

Constraint Manifold RL[51] focuses on safe exploration through constraint manifold methods in robotics, while the original paper addresses safe exploration via policy priors with different theoretical guarantees and algorithmic approaches.

2. Reinforcement Learning Control of Shape Memory Alloy Based Soft Robotic Platform

URL: View paper

Brief Assessment

Shape Memory Alloy Control[57] focuses on soft robotics with shape memory alloy actuators, not safe RL algorithms. The hardware validation is for a completely different domain (soft robotic limbs) rather than safe exploration in general RL tasks.

3. Safe learning in robotics: From learning-based control to safe reinforcement learning

URL: View paper

Brief Assessment

Safe Learning Robotics[55] is a survey paper reviewing safe learning methods in robotics broadly, not presenting a specific algorithm with empirical validation. It does not demonstrate a novel safe RL algorithm on benchmarks and hardware that would refute SOOPER's novelty claim.

4. Bresa: Bio-inspired Reflexive Safe Reinforcement Learning for Contact-Rich Robotic Tasks

URL: View paper

Brief Assessment

Bio-Inspired Reflexive Safe RL[58] focuses on contact-rich robotic tasks with a hierarchical safety critic operating at low-level control frequencies, distinct from SOOPER's model-based safe exploration framework with policy priors. The candidate validates on contact-rich manipulation tasks, while the original validates on safe RL benchmarks (RWRL, SafetyGym) and different hardware platforms.

5. Recovery rl: Safe reinforcement learning with learned recovery zones

URL: View paper

Brief Assessment

Recovery RL Zones[50] focuses on safe exploration using recovery policies for navigation and manipulation tasks, not on general safe RL frameworks with policy priors like SOOPER. The candidate's hardware validation is specific to obstacle avoidance rather than the broader benchmark evaluation claimed by the original paper.

6. Multi-robot hierarchical safe reinforcement learning autonomous decision-making strategy based on uniformly ultimate boundedness constraints

URL: View paper

Brief Assessment

Multi-Robot Hierarchical Safe RL[54] focuses on multi-robot systems with uniformly ultimate boundedness constraints and validates on simulation benchmarks (Safety-Gymnasium, CoppeliaSim) but does not report real-world hardware experiments. The original paper validates SOOPER on both standard safe RL benchmarks and real-world robotic hardware (race car), which is a distinct contribution.

7. A review of safe reinforcement learning: Methods, theories and applications

URL: View paper

Brief Assessment

Safe RL Review[49] is a survey paper that reviews existing safe RL methods and their applications. It does not present a novel algorithm with empirical validation that could refute SOOPER's novelty claim of being the first to demonstrate its specific approach on benchmarks and real hardware.

8. Benchmarking actor-critic deep reinforcement learning algorithms for robotics control with action constraints

URL: View paper

Brief Assessment

Benchmarking Actor-Critic Robotics[53] focuses on benchmarking actor-critic algorithms for robotics control with action constraints in simulation environments. It does not address safe exploration in reinforcement learning with policy priors or validate approaches on real-world hardware for safe RL as the original paper does.

9. Safety gymnasium: A unified safe reinforcement learning benchmark

URL: View paper

Brief Assessment

Safety Gymnasium Benchmark[52] focuses on providing a unified benchmark suite for safe RL evaluation, not on proposing a specific algorithm with theoretical guarantees validated on hardware. The candidate presents environments for testing, while the original paper proposes SOOPER, a novel algorithm with safety guarantees.

10. Adaptive Safety-Certified Reinforcement Learning for Constrained Optimal Control of Autonomous Robots With Uncertainties

URL: View paper

Brief Assessment

Adaptive Safety-Certified RL[56] focuses on control barrier functions for constrained optimal control with uncertainties, not on safe RL benchmarks or the specific hardware validation approach described in the original paper.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Safe Exploration via Policy Priors View paper
- [1] Real-Time Optimal Power Flow Method via Safe Deep Reinforcement Learning Based on Primal-Dual and Prior Knowledge Guidance View paper
- [2] Bayesian controller fusion: Leveraging control priors in deep reinforcement learning for robotics View paper
- [3] Safe exploration for interactive machine learning View paper
- [4] Quantum computing and neuromorphic computing for safe, reliable, and explainable multi-agent reinforcement learning: optimal control in autonomous robotics View paper
- [5] Provably good batch off-policy reinforcement learning without great exploration View paper
- [6] Safe Adaptive Policy Transfer Reinforcement Learning for Distributed Multiagent Control View paper
- [7] Enhancing efficiency of safe reinforcement learning via sample manipulation View paper
- [8] Reinforcement learning with adaptive regularization for safe control of critical systems View paper
- [9] Trajectory-wise iterative reinforcement learning framework for auto-bidding View paper
- [10] Learning transferable domain priors for safe exploration in reinforcement learning View paper
- [11] Jump-Start Reinforcement Learning with Self-Evolving Priors for Extreme Monopedal Locomotion View paper
- [12] Specialized Deep Residual Policy Safe Reinforcement Learning-Based Controller for Complex and Continuous State-Action Spaces View paper
- [13] Safe Exploration Method for Reinforcement Learning under Existence of Disturbance View paper
- [14] Rocket Landing Control with Random Annealing Jump Start Reinforcement Learning View paper
- [15] Coordination of Bounded Rational Drones Through Informed Prior Policy View paper
- [16] Reliable Reinforcement Learning for Decision-Making in Autonomous Driving View paper
- [17] Planning with RL and episodic-memory behavioral priors View paper
- [18] Multiplicative Controller Fusion: Leveraging Algorithmic Priors for Sample-efficient Reinforcement Learning and Safe Sim-To-Real Transfer View paper
- [19] Learning safe policies with expert guidance View paper
- [20] Training and Evaluation of Deep Policies using Reinforcement Learning and Generative Models View paper
- [21] Accelerating Safe Reinforcement Learning with Constraint-mismatched Policies View paper
- [22] Anytime-competitive reinforcement learning with policy prior View paper
- [23] Data-efficient visuomotor policy training using reinforcement learning and generative models View paper
- [24] Safe exploration in reinforcement learning: Theory and applications in robotics View paper
- [25] Accelerating Safe Reinforcement Learning with Constraint-mismatched Baseline Policies View paper
- [26] Safetyâ□□Efficiency Balanced Navigation for Unmanned Tracked Vehicles in Uneven Terrain Using Prior-Based Ensemble Deep Reinforcement Learning View paper
- [27] Probabilistic policy reuse for safe reinforcement learning View paper
- [28] Amortized Safe Active Learning for Real-Time Data Acquisition: Pretrained Neural Policies from Simulated Nonparametric Functions View paper
- [29] TempoRL: Temporal Priors for Exploration in Off-Policy Reinforcement Learning View paper
- [30] Training and Transferring Safe Policies in Reinforcement Learning View paper
- [31] Constrained Residual Race: An Efficient Hybrid Controller for Autonomous Racing View paper
- [32] Learning to navigate through abstraction and adaptation View paper
- [33] Safe Exploration Incurs Nearly No Additional Sample Complexity for Reward-free RL View paper
- [34] Directed Policy Gradient for Safe Reinforcement Learning with Human Advice View paper
- [35] From Game-Playing to Self-Driving: Comparing AlphaGo vs AlphaZero Approaches for Driving Controls View paper
- [36] Safe model-based reinforcement learning with stability guarantees View paper
- [37] A KL-regularization framework for learning to plan with adaptive priors View paper

- [38] Context-Aware Policy-Guided Gradient Search for Offline Model-Based Optimization View paper
- [39] Safe reinforcement learning in constrained markov decision processes View paper
- [40] Safe reinforcement learning with contextual information: Theory and applications View paper
- [41] Rethinking safe policy learning for complex constraints satisfaction: A glimpse in real-time security constrained economic dispatch integrating energy storage units View paper
- [42] Adversarially Trained Weighted Actor-Critic for Safe Offline Reinforcement Learning View paper
- [43] Conservative safety critics for exploration View paper
- [44] Probabilistic constraint for safety-critical reinforcement learning View paper
- [45] Truly no-regret learning in constrained mdps View paper
- [46] Safety and robustness in reinforcement learning View paper
- [47] Regret guarantees for model-based reinforcement learning with long-term average constraints View paper
- [48] Triple-q: A model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation View paper
- [49] A review of safe reinforcement learning: Methods, theories and applications View paper
- [50] Recovery rl: Safe reinforcement learning with learned recovery zones View paper
- [51] Robot reinforcement learning on the constraint manifold View paper
- [52] Safety gymnasium: A unified safe reinforcement learning benchmark View paper
- [53] Benchmarking actor-critic deep reinforcement learning algorithms for robotics control with action constraints View paper
- [54] Multi-robot hierarchical safe reinforcement learning autonomous decision-making strategy based on uniformly ultimate boundedness constraints View paper
- [55] Safe learning in robotics: From learning-based control to safe reinforcement learning View paper
- [56] Adaptive Safety-Certified Reinforcement Learning for Constrained Optimal Control of Autonomous Robots With Uncertainties View paper
- [57] Reinforcement Learning Control of Shape Memory Alloy Based Soft Robotic Platform View paper
- [58] Bresa: Bio-inspired Reflexive Safe Reinforcement Learning for Contact-Rich Robotic Tasks View paper