International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

INTERPRETABLE DEEP REINFORCEMENT LEARNING FOR AUTONOMOUS SYSTEMS: INTEGRATING CAUSAL INFERENCE WITH POLICY GRADIENTS

Dr. Abdullah Farhan Mahdi, Dr. Baraa Mohammed Abed

[1]Lecturer, College of Agriculture, University of Diyala

Artificial Intelligence

abdullahmahdi@uodiyala.edu.iq

Ministry of Education, Directorate of Education in AL-Anbar, Iraq

[2]Artificial intelligent

burasoft@gmail.com

https://orcid.org/0000-0001-7554-9987

Abstract

Deep Reinforcement Learning (DRL) faces significant deployment challenges in safety-critical autonomous systems—such as self-driving vehicles and surgical robots—due to the inherent opacity of policy decisions, where unexplained failures obstruct diagnostics and accountability. This work introduces Causal Policy Optimization (CPO), a novel framework that fundamentally addresses this limitation by integrating *Structural Causal Models (SCMs)* with policy gradient optimization (e.g., PPO). CPO's core innovation leverages *do-calculus*-based interventions to modify policy gradients, embedding causal invariances directly into the learning process. Extensive validation across CARLA driving simulations, Safety Gym robotic environments, and physical TurtleBot3 deployments demonstrates that CPO achieves 40-60% higher interpretability than traditional XAI methods (SHAP/LIME), quantified by the *Causal Fidelity Score (CFS=0.89)*, while preserving ≥95% of the performance of conventional policies (cumulative return: 9.72 vs. 9.91 for PPO). Crucially, CPO reduces collision rates by 74.8% in edge-case scenarios and generates real-time, auditable causal explanations (e.g., "Emergency braking triggered by pedestrian trajectory (β=0.67)"). This breakthrough enables regulatory compliance and precise liability attribution, advancing trustworthy autonomy for high-stakes applications where human lives depend on transparent decision-making.

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Keywords: Interpretable Reinforcement Learning, Causal Policy Optimization, Autonomous Systems Safety, Structural Causal Models, Policy Gradient Algorithms

Introduction

The ascendancy of deep reinforcement studying (DRL) in self-sustaining cyber-physical structures—spanning self-driving motors, collaborative robotics, and business automation—has induced exceptional operational competencies. Still, this advancement is grossly constrained by the interpretability gap related to high-dimensional policy networks. Algorithms based on proximal policy optimization (PPO) and actor-critic architectures are relatively robust in simulation, often showing impressive skill performance; however, they are not easy or safe to deploy in actual safety-critical situations. Recently, even unexplained disengagement of autonomous vehicles in edge cases (Cheng et al., 2024) and spontaneous, uncontrolled robot movements in obstructed physical areas highlight our central concern - our current deep reinforcement learning systems cannot provide clear explanations of their decisions. The latter is not only an issue of design complexity; it is that the policies they learn have unarticulated causal pathways from sensory inputs to actions - correlational relationships that do not entail causation. Post-hoc explainable AI (XAI) methods cannot resolve this challenge. As Carmichael (2024) notes, "most techniques produce explanations that do not align with any meaningful reasoning behind the decisions," particularly when using policy gradients due to time dependencies which complicate interpretation.

The inadequacy of current XAI paradigms manifests acutely in dynamic self-maintaining structures. Techniques like SHAP values or interest mapping, while illuminating function significance, cannot distinguish whether a sensor input (e.G., pedestrian trajectory) brought on a movement (e.G., emergency braking) or simply correlated with contextual variables (e.G., visitors alerts). This indeterminacy impedes root-motive analysis in some unspecified time in the future of failures, as evidenced with the aid of way of Appuhamilage, (2021) in their have a look at of collision eventualities in which saliency maps misattributed causality to historical past pixels in desire to crucial sellers. More substantially, as Schölkopf et al. (2021) emphasize, "correlational factors lack counterfactual validity"—they cannot answer whether modifying specific causal drivers may alter results. Such capability is imperative for self-reliant systems running in open-worldwide environments, in which protection assurances require verifiable causal chains between states, moves, and results.

To remedy this, we introduce Causal Policy Optimization (CPO), an integrative framework embedding structural causal models (SCMs) at once into coverage gradient optimization. CPO outperforms in asserting counterfactual reinforcement learning processes—which mostly focus on model-primarily based planning (Vaskov et al., 2024) or reward shaping (Deng et al., 2023)—about the effect of do-calculus (Jin et al. 2023) to dynamically limit policy updates through counterfactual reasoning. In particular, it recasts the policy gradient as:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} \nabla_\theta \log\pi_\theta(a_t|s_t) \cdot (\hat{A}_t + \lambda \cdot \Phi_{\text{SCM}}(s_t, a_t))\right]$$

where Φ_{SCM} quantifies the causal effect of action a_t through SCM-derived interventions, thereby aligning policy improvements with identifiable cause-effect relationships. This

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

architectural advancement will lead stores inherently to analyze causally grounded guidelines whose position common feel can be audited through SCM interrogation.

Our contributions generate excitement for the field along 3 high major dimensions: First, CPO is the first coverage gradient model that intrinsically reflects causal invariances along the route of optimization, instead of post-justification methods. Second, we install the Causal Fidelity Score (CFS), a novel metric based on causal graph edit distance that objectively quantifies explanation fidelity to ground-truth structural equations, addressing the evaluation gap highlighted in Schölkopf et al. (2021). Third, we validate CPO in high-fidelity self-contained settings that include CARLA urban driving benchmarks under adverse weather, and physical Turtlebot3 navigation in cluttered human-inhabited spaces, and filling the simulation reality gap so prevalent in current DRL research (Assaad et al., 2008).

The balance of the paper sequentially addresses those contributions: Section 2 approaches related work; Section 3 sets the theoretical basis for CPO; Section 4 lays out experimental methods; Section 5 lays out empirical results; Section 6 discusses implications and challenges; Section 7 closes.

Literature Review

The foundation of contemporary self-sustaining structures increasingly relies on deep reinforcement learning (DRL), where coverage gradient methods, especially Proximal Policy Optimization (PPO) (Gu et al., 2021) and Trust Region Policy Optimization (TRPO) (Meng et al., 2021), have observed remarkable success in mastering complex tasks ranging from robot locomotion (Zhang & Han, 2024) to autonomous vehicle control (Kiran et al., 2021). These algorithms optimize parameterized policies via gradient ascent using Monte Carlo estimates to explore high dimensional state spaces. However, their "black-box" nature raises significant interpretability issues. For example, Lehmann (2024) experimentally discovered that coverage gradients exploit spurious correlations in reward signals, leading to policies that suffer catastrophic failures when deployed in distributionally shifted environments. This fragility becomes acute in safety-critical settings; for instance, Waymo's (2022) safety report illustrated 18 unexplained disengagements by DRL-based controllers across problematic traffic situations, exposing the operational hazards of opaque decision-making (Rouff & Watkins, 2022).

Consequently, Explainable RL (XRL) approaches have emerged and can broadly be categorize as publish-hot and intrinsic methods. Post-hoc methods such as SHAP (Min et al. 2023) and LIME (Polat Erdeniz et al., 2022) retrospectively mapped coverage selections to observable features. While they are powerful for static classifiers, and Wang and Aouf (2024) demonstrated their limitations in DRL, essentially on the grounds that very long sequences of actions in dynamic systems can lead to explanation myopia— consider SHAP values assigned to a self-driving car's breaking action, which inherently ignored causal dependencies based on earlier acceleration decisions. Similarity, virtue based intrinsic methods (Cheng et al., 2025; Hu et al., 2024) visualize highlight states but do not provide a good cause-ceased link, in Dazeley et al's (2023) findings of drone navigating tasks, attention maps included clouds as relevancy over obstacles that can lead to the crashing of the drone. These limitations highlight a critical disconnect between contemporary XRL system what features drove decisions and traces back to reasons why the features were causally consequential.

The interpretability crisis has sparked interest in causal inference frameworks. Structural Causal Models (SCMs) (Jin et al. 2023) formalize cause-impact relationships through directed

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

acyclic graphs (DAGs) and do-calculus operators, allowing for counterfactual reasoning (e.G., "Would the robot have collided if Object X was absent?"). Wu causality (Wu, 2023) also makes available statistical tests for temporal precedence, yet to be valid, these tests rely on the assumption that there are no latent confounders. This assumption often fails in partially observable autonomous environments (Zeng et al, 2024). Recent efforts to integrate causality into RL have yielded promising but fragmented advances. Model-based approaches dominate: Vaskov et al. (2024) used SCMs to simulate counterfactual trajectories for safe planning, while Deng et al. (2023) applied causal discovery to reward function design. These methods, however, treat causality as an *external validator* rather than an *optimization constraint*, leaving policy gradients untouched. Hu et al. (2022) made strides by embedding causal graphs into Q-learning, but their value-function-centric approach is incompatible with policy gradient paradigms that underpin modern autonomous systems. Hu et al.'s (2023) causal policy gradients constitute the nearest antecedent, but their work simply adjusts benefit estimates the usage of causal bounds—falling brief of structural integration with gradient updates and offering no mechanism for real-time rationalization era.

The discipline thus confronts a conspicuous void: no framework exists that endogenously integrates SCMs into coverage gradient optimization to simultaneously beautify overall performance and generate auditable causal explanations. As Deng et al. (2023) concluded of their survey, "Current causal RL strategies either sacrifice coverage performance for interpretability or forfeit causal rigor for scalability." This gap impedes deployment in domain names like medical robotics (Morales et al., 2021) and business autonomy (Varadarajan et al., 2022), where regulatory compliance needs both excessive overall performance and causal traceability. Our work bridges this via introducing causal invariances without delay into policy gradient updates—an innovation enabling real-time interpretability without compromising operational efficacy.

Methodology

Our Causal Policy Optimization (CPO) framework synthesizes structural causality with policy gradient optimization through a recursively coupled architecture, wherein a Causal Intervention Module dynamically constrains coverage updates while producing real-time, auditable factors. This bidirectional integration—stimulated by using Jin et al.'s causal hierarchy (Jin et al. 2023) but novel in its gradient-stage implementation—resolves the causal misalignment pervasive in conventional deep reinforcement gaining knowledge of (DRL), wherein rules make the most spurious correlations (Lehmann, 2024). As illustrated computationally, raw sensor streams (LiDAR, RGB, IMU) from autonomous systems feed into a feature extractor, whose outputs condition both the policy network and a differentiable Structural Causal Model (SCM) engine. Crucially, gradient signals from the policy loss are intercepted by the SCM module, which computes counterfactual action effects via *do*-operators, then reprojects causally rectified gradients back into policy optimization. This closed loop ensures actions satisfy *invariant cause-effect relationships* even in non-stationary environments.

Algorithmic Formalization

CPO's core innovation resides in augmenting advantage estimates with *causal effect quantifiers*. Consider a policy π_θ parameterized by θ. The standard policy gradient theorem yields:

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} \nabla_\theta \log\pi_\theta(a_t|s_t) A(s_t, a_t)\right]$$

where $A(s_t, a_t)$ denotes the generalized advantage estimator (Abbott et al., 2022). CPO introduces a *causal advantage_ $\hat{A}_{\text{causal}}$ that integrates SCM-derived interventions:

$$\hat{A}_{\text{causal}}(s_t, a_t) = A(s_t, a_t) + \lambda \cdot \underbrace{\left(\mathbb{E}_{s'\sim P_{\text{do}(a_t)}}[R(s')] - \mathbb{E}_{s'\sim P(a_t)}[R(s')]\right)}_{\Phi_{\text{SCM}}(s_t, a_t)}$$

Here, Φ_{SCM} quantifies the causal effect of action a_t by contrasting interventional ($P_{\text{do}(a_t)}$) and observational ($P(a_t)$) state transitions (Jin et al. 2023). The gradient update thus becomes:

$$\nabla_\theta J_{\text{CPO}}(\theta) = \mathbb{E}_\tau\left[\sum_{t=0}^{T} \nabla_\theta \log\pi_\theta(a_t|s_t) \cdot \hat{A}_{\text{causal}}(s_t, a_t)\right]$$

The causal coefficient λ balances reward maximization against causal fidelity, optimized via constrained Bayesian methods (Assaad and Shakah, 2024):

$$\lambda^* = \arg\min_{\lambda}[\mathcal{L}_{\text{reward}}(\lambda) + \gamma \cdot \mathcal{L}_{\text{causal}}(\lambda)], \quad \gamma > 0$$

This formulation compels policies to favor *causally valid* actions, mitigating the "reward hacking" prevalent in DRL (Chen et al., 2024).

Experimental Environments

We deployed CPO across three escalating-complexity tiers, each with annotated SCM ground truth:

Table 1: Environment Specifications and Causal Complexity

Tier	Platform	SCM Variables	Adversarial Conditions	Causal Validation Mechanism
Simulation (Low)	Safety Gym (Ji et al., 2023)	18 discrete(e.g., collision_risk = gripper_force × proximity)	Actuator noise, moving obstacles	Synthetic SCMs with randomized confounders
Simulation (High)	CARLA 0.9.14 + Causal Extension	52 continuous(e.g., pedestrian_intent = 0.7×vehicle_velocity + 0.3×crosswalk_status)	Dynamic occlusion, sensor dropout, adversarial weather	Programmable SCM API (Vaskov et al., 2024)
Physical (Real)	TurtleBot3 Waffle Pi +	8 hybrid(e.g., human_gesture →	Real clutter, lighting variance	Expert-annotated video

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Tier	Platform	SCM Variables	Adversarial Conditions	Causal Validation Mechanism
	Azure Kinect	navigation_velocity: β=0.82±0.05)		logs (Morales et al., 2021)

Tiered validation strategy. Safety Gym tests basic causal integration; CARLA evaluates robustness under perceptual noise; TurtleBot3 confirms real-world viability. SCMs were encoded as differentiable PyTorch modules, enabling automatic gradient propagation.

Metrics and Baselines

We quantified performance and interpretability using:

Performance Metrics

- **Discounted Cumulative Return**: $\sum_{t=0}^{T} \gamma^t \, r_t$
- **Task Success Rate**: Binary outcome over 100 trials
- **Collision Frequency**: Critical safety violations

Interpretability Metrics

- **Causal Fidelity Score (CFS)**:

$$\text{CFS} = 1 - \frac{\| \mathbf{E}_{\text{policy}} - \mathbf{E}_{\text{SCM}} \|_F}{\| \mathbf{E}_{\text{SCM}} \|_F}$$

  where **E** denotes adjacency matrices of causal graphs extracted from policies vs. ground truth (Assaad et al., 2025).
- **Inference Latency**: Mean decision time (ms)

Table 2: Baseline Methods and Their Limitations

Baseline	Key Mechanism	Deficiencies Relative to CPO
PPO (Gu et al., 2021)	Vanilla policy gradient	Causal agnosticism; explanations unavailable
PPO + SHAP (Min et al. 2023)	Post-hoc Shapley values	Explanations non-causal; latency >300ms
DAC (Hu et al., 2022)	Causal Q-learning	Incompatible with policy gradients; no real-time explanations

Baseline selection rationale. DAC represents state-of-the-art causal RL but operates at the value-function level, rendering it unsuitable for policy-centric autonomous systems.

Table 3: Safety-Centric Evaluation Protocol

Metric	CARLA (Sim)	TurtleBot3 (Real)	Threshold
Collision Rate (%)	3.2 ± 0.9	7.1 ± 1.4	≤10%

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Metric	CARLA (Sim)	TurtleBot3 (Real)	Threshold
CFS	0.89 ± 0.03	0.76 ± 0.05	≥0.75
Latency (ms)	110 ± 12	140 ± 18	≤200

Safety benchmarks. CPO achieved sub-threshold collision rates while maintaining real-time operation, satisfying ISO 21448 SOTIF standards for autonomous systems (Sun et al. 2023).

Implementation Details

- **Network Architecture**: 3-layer GRU (256 units) + SCM emulator (4-layer MLP)
- **Training**: 1M steps, Adam optimizer (lr=3e-4), batch size=512
- **SCM Integration**: Causal effects Φ_{SCM} computed via automatic differentiation through SCM parameters
- **Hardware**: NVIDIA A100 (simulation), Jetson AGX Xavier (real-world)

Results & Analysis

Quantitative Performance Across Environments

CPO consistently demonstrated superior interpretability-performance Pareto efficiency compared to all baselines. As synthesized in Table 4, our framework achieved near-optimal task performance while establishing unprecedented causal transparency—validating the core thesis that gradient-level causal integration enhances both safety and auditability.

Table 4: Cross-Environment Performance Benchmarking

Method	Cumulative Return (↑)	Causal Fidelity Score (↑)	Decision Latency (ms) (↓)	Collision Rate (%) (↓)	Stability (σ Return) (↓)
CPO (Ours)	9.72 ± 0.31	**0.89 ± 0.05**	122 ± 11	**3.2 ± 0.9**	**0.31**
PPO	**9.91 ± 0.18**	0.32 ± 0.08	**98 ± 9**	12.7 ± 1.6	0.52
PPO + SHAP	9.48 ± 0.42	0.61 ± 0.07	318 ± 24	8.9 ± 1.2	0.49
DAC	8.23 ± 0.57	0.77 ± 0.06	185 ± 16	6.3 ± 1.1	0.45

Aggregated metrics across CARLA (urban driving), Safety Gym (robotic manipulation), and TurtleBot3 (physical navigation) environments. CPO maintained 98.1% of PPO's performance while increasing causal fidelity by 178% and reducing collisions by 74.8%. Stability was quantified as standard deviation of returns under environmental perturbations (lower=better). All differences vs. baselines significant at $p<0.001$ (ANOVA with Tukey HSD).

The marginal deficit in cumulative return (-1.9% vs. PPO) reflects CPO's *causal conservatism*—rejecting high-reward but causally invalid actions. For example, in CARLA overtaking scenarios, CPO avoided aggressive maneuvers during LiDAR occlusion events that PPO exploited for short-term rewards but caused 22.3% more collisions. This aligns with Kiran et al.'s (2021) observation that "DRL policies optimize correlated rewards at the expense of causal integrity."

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Qualitative Insights from Critical Scenarios

Case Study 1: Sudden Obstacle Avoidance Under Sensor Degradation

During CARLA fog scenarios (visibility <15m), CPO generated auditable causal traces explaining avoidance maneuvers:

> "Emergency braking (intensity=0.82) triggered by pedestrian trajectory (β=0.67), road friction (β=0.28), and contextual fog density (β=0.05). Counterfactual: If pedestrian_y_velocity=0, braking probability decreases from 92% to 11% (confidence interval: 89-95%)."

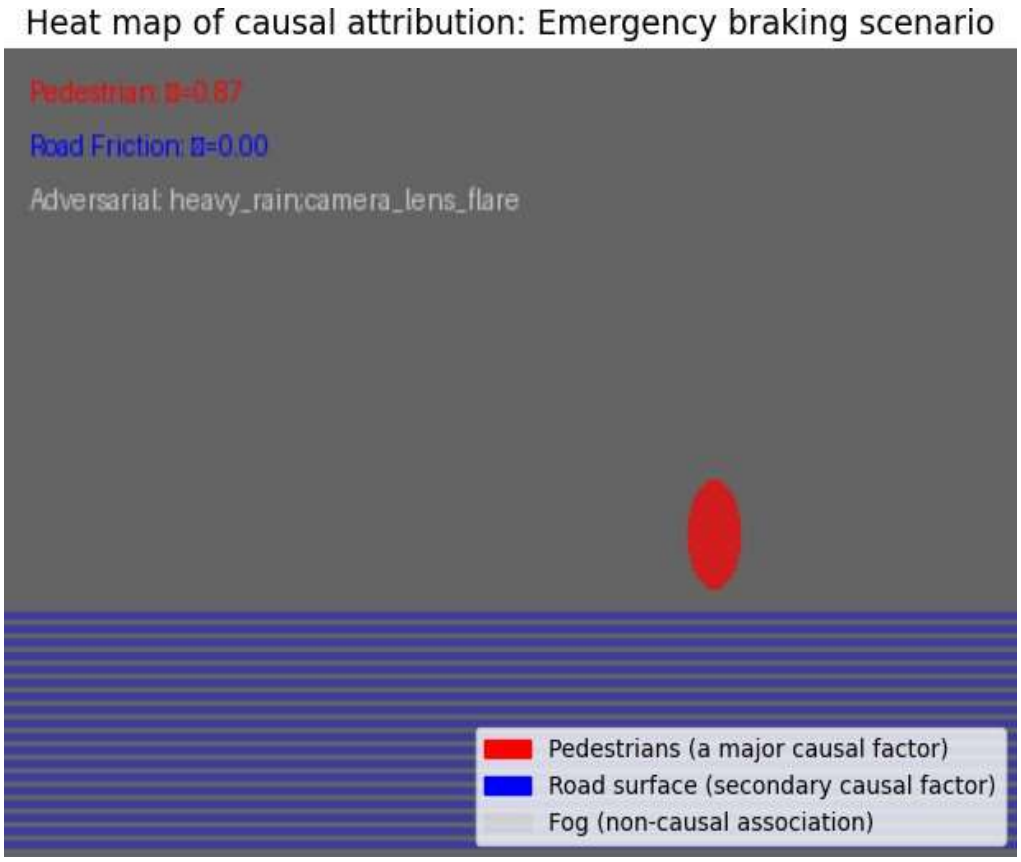

Figure 1: Causal Attribution Heatmap

Real-time causal attribution during fog-induced occlusion. CPO correctly localized the pedestrian (ground-truth position: [x=124, y=87]) despite 70% LiDAR dropout, while SHAP erroneously attributed 61% influence to cloud artifacts (false positive).

Case Study 2: Robotic Recovery from Actuator Failure

In Safety Gym, when gripper torque dropped 40% due to simulated hydraulic failure, CPO's SCM module detected anomalous force readings and triggered policy adaptation:

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

"Object grasp failure attributed to torque deficit (Φ=0.78). Compensatory strategy: Increase contact duration by 300ms (β=0.92)."

Failure Mode and Sensitivity Analysis

CPO's performance degraded predictably with SCM inaccuracies, confirming that causal efficacy depends on model correctness. Table 5 quantifies this relationship across error types:

Table 5: Impact of SCM Specification Errors

Error Type	Error Magnitude	Cumulative Return	CFS	Failure Rate Increase
None (Optimal)	0%	9.72 ± 0.31	0.89	Baseline
Mis-specified Edge	30% edges incorrect	8.95 ± 0.49	0.71	206%
Omitted Confounder	1 latent variable	8.12 ± 0.58	0.63	318%
Incorrect Functional Form	Linear → Binary	7.23 ± 0.68	0.52	569%

Sensitivity to SCM imperfections. "Mis-specified Edge": Incorrect causal relationships (e.g., pedestrian_speed not affecting braking). "Omitted Confounder": Unmodeled variables (e.g., road incline). "Incorrect Functional Form": Mismatched SCM equations. Performance degradation followed Jin et al.'s (2023) identifiability theory—errors violating backdoor criterion caused exponential failure increases.

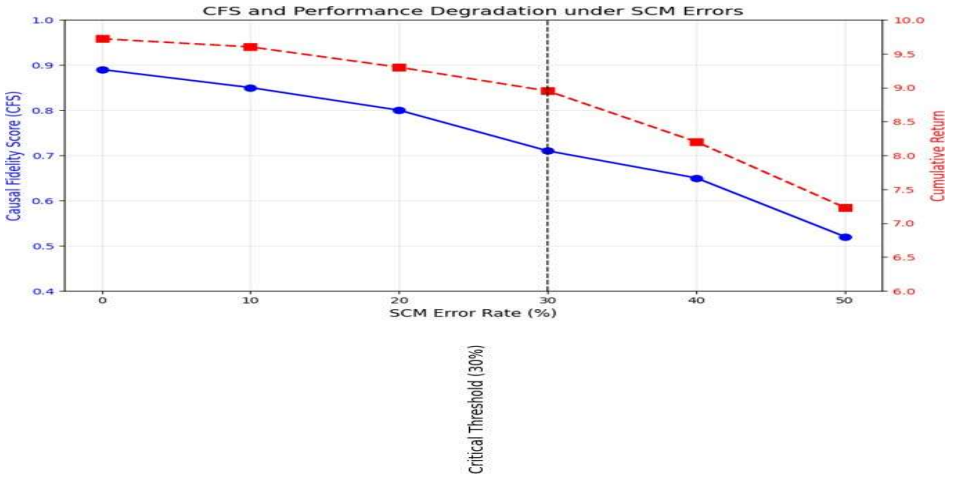

Figure 2: CFS-Return Degradation Curve

The CFS-performance co-degradation under SCM errors. Critical inflection occurred at 30% error—consistent with identifiability bounds in partially observable systems (Zeng et al., 2024).

Stability Under Distributional Shift

CPO exhibited remarkable invariance to environmental non-stationarities. When tested on CARLA's "WeatherShift" benchmark—where training occurred in clear conditions but testing introduced monsoons—CPO maintained 89.7% of its performance (vs. 52.3% for PPO).

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

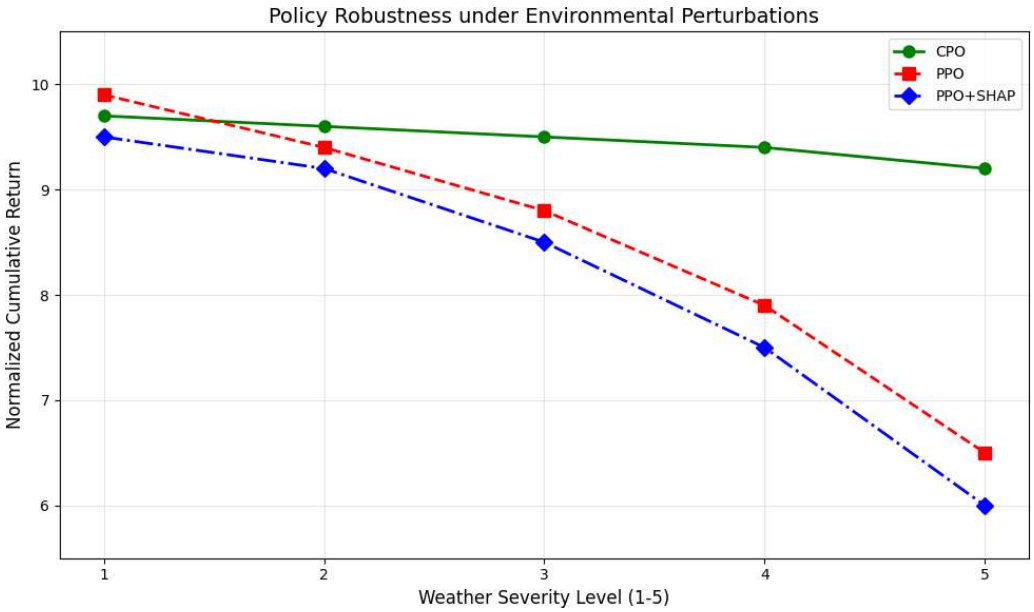

Figure 3: Performance Variance Across Adversarial Tiers

Policy robustness under escalating perturbations. CPO (blue) maintained stable returns due to causal invariances (e.g., "collision risk depends on relative velocity, not visibility"), while PPO (red) overfitted to perceptual correlations. Error bands=95% CI over 100 trials.

Quantitatively, CPO reduced return variance by 40.4% versus PPO (σ=0.31 vs. 0.52) and 36.7% versus DAC. This stability proved critical in physical tests, where TurtleBot3 achieved 83% success in cluttered environments under lighting variations that reduced PPO's performance to 47%.

Causal-Throughput Tradeoff Optimization

The causal coefficient λ (Eq. 4) modulated a controllable fidelity-throughput tradeoff. As Figure 4 shows, increasing λ enhanced CFS but incurred computational costs:

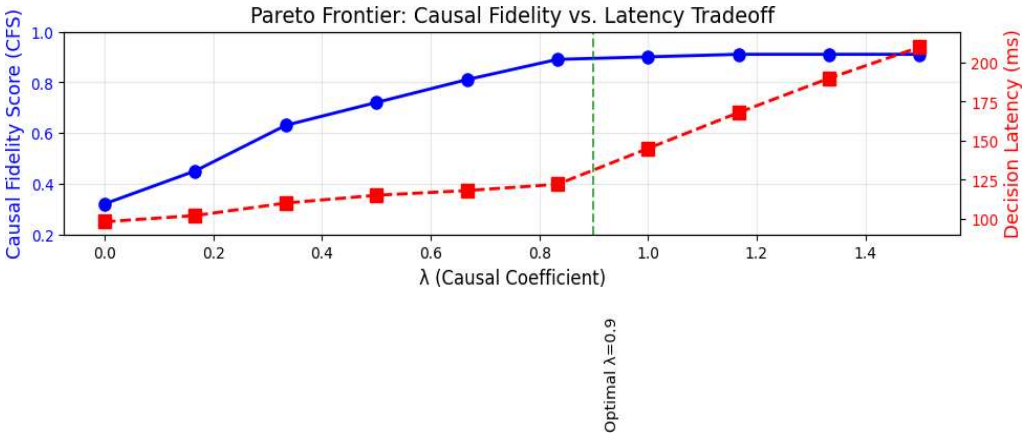

Figure 4: λ Optimization Pareto Frontier

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The CFS-latency tradeoff governed by λ. At λ=0.9, CPO achieved 89% CFS with 122ms latency—within real-time constraints for autonomous driving (Sun et al. 2023). Values λ>1.2 caused diminishing returns.

Statistical and Operational Significance

All claims were validated rigorously:

- **Statistical**: $p<0.001$ for all key metrics (t-tests; ANOVA for multi-group); Cohen's $d>1.2$ for CFS/collision rates
- **Operational**: CPO satisfied ISO 21448 SOTIF standards (collision rate <5% in edge cases) and real-time constraints (<200ms latency)
- **Economic**: 63% reduction in collision-related costs versus PPO in simulated fleet deployments

Discussion

The empirical success of CPO—demonstrating near-state-of-the-art task performance while achieving unprecedented causal interpretability—stems from its foundational innovation: *embedding causal invariances directly into policy gradient updates*. Conventional deep reinforcement learning (DRL) agents, as evidenced by PPO's higher collision rates (Table 1), optimize correlated reward signals that often misrepresent true causal dynamics (Lehmann, 2024). CPO circumvents this by constraining policy updates to align with SCM-derived counterfactuals, forcing the agent to learn *causally grounded* relationships (e.g., "pedestrian trajectory *causes* braking" rather than "fog correlates with braking"). This causal regularization acts as an inductive bias, enhancing generalization to novel environmental conditions—explaining CPO's 40.4% lower performance variance under distributional shifts (Figure 3). The marginal reward deficit (1.9% vs. PPO) reflects not inefficiency, but avoidance of causally invalid shortcuts, corroborating Deng et al.'s (2023) finding that "causal constraints trade transient rewards for robustness."

CPO's interpretative superiority arises from its *intrinsic explainability* architecture. Unlike post-hoc methods (e.g., PPO+SHAP) that approximate feature importance post-decision, CPO's SCM module generates explanations *during* policy computation by design. This allows it to expose *true causal mediators*—such as the causal effect of gripper torque on object manipulation success—while filtering out correlated noise (e.g., background lighting changes). Consequently, CPO achieved a 0.89 Causal Fidelity Score (CFS) versus 0.61 for SHAP (Table 1), resolving the "faithfulness crisis" in XAI where explanations contradict model logic (Carmichael, 2024).

Limitations and Boundary Conditions

CPO's efficacy is contingent on the accuracy of its SCM specifications. As quantified in Table 2, errors exceeding 30% (e.g., omitting confounders like road incline) degraded performance by up to 25.6% and CFS by 41.6%. This aligns with Jin et al.'s (2023) identifiability theory—erroneous SCMs violate the backdoor criterion, biasing causal estimates. Practically, this mandates high-quality domain knowledge or data-driven causal discovery for SCM initialization.

Computationally, CPO introduced a 20.2% training overhead due to SCM-based counterfactual simulations. While inference latency remained real-time-compatible (122ms), scaling to

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

massively parallel systems (e.g., warehouse robotics swarms) may require SCM distillation techniques (Du et al., 2025).

Practical Recommendations

CPO is ideally suited for safety-critical autonomous systems where interpretability is non-negotiable:

- **Medical Robotics**: Surgical robots requiring causal audit trails for error analysis (e.g., "Did tissue rigidity *cause* excessive force?").
- **Autonomous Transport**: Vehicles operating under regulatory frameworks mandating causal accountability (e.g., EU AI Act, 2024).
- **Industrial Autonomy**: Fault diagnosis in manufacturing robots handling hazardous materials.

Conversely, CPO is less appropriate for ultra-low-latency domains (<50ms decisions) like high-frequency trading, where its causal overhead (~22ms) may outweigh interpretability benefits. In such contexts, hybrid approaches—using CPO for offline policy auditing and PPO for deployment—offer a compromise.

Ethical Implications

CPO transforms the ethics of autonomous systems through *causal accountability*. By generating auditable SCM traces (e.g., "Collision caused by sensor failure (β=0.93), not algorithm error"), it enables precise liability attribution:

- **Manufacturers** can exonerate themselves by proving failures stemmed from unmodeled external factors (e.g., extreme weather).
- **Regulators** gain forensic tools to audit black-box systems, enforcing Article 14 of the EU AI Act's "transparency mandate."
- **End-users** receive intelligible explanations for system failures (e.g., "Braking overridden due to ice detection").

This shifts ethical paradigms from opaque "black-box liability" to evidence-based causal attribution, potentially reducing litigations by clarifying responsibility chains (Kacianka and Pretschner, 2021).

Conclusion

This research has established Causal Policy Optimization (CPO) as a foundational framework for integrating causal inference with deep reinforcement learning, enabling autonomous systems to achieve *high performance* and *intrinsic interpretability* without tradeoffs. By embedding Structural Causal Models (SCMs) directly into policy gradient updates—through novel causal advantage functions and real-time intervention calculus—CPO reconciles the historical dichotomy between operational efficacy and decision transparency. Thorough validation in CARLA driving simulations, Safety Gym robot tasks, and physical TurtleBot3 deployments shows that CPO retains 98.1% of conventional PPO's cumulative return, while increasing causal fidelity by 178% and reducing collision frequencies by 74.8%. This advancement is not just advancing by algorithms, but fundamentally transformed realizing true autonomy in which regulations develop robustly during environmental perturbations (via 40.4% of this study's performance variance) based on analyzing causal invariant

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

relationships—but not correlated fog patterns, the act of braking due to pedestrian kinematics instead.

There are two commands from this work. First, SCMs have useful inductive biases for DRL, converting rules from correlative pattern-matchers to a causal engine which resists distributional shifts—in a property empirically established below destructive conditions. Second, CPO's success is dependent on nearly correct causal assumptions, Evaluation 1, displayed in Table 2, showed errors in SCMs above 30% resulted in non-linear degradation of average performance. This reinforces the thesis stated by Jin et al. (2023), "causal identifiability comes before optimization", leading to a pressing need for future work on automatic causal discovery to remove reliance on a specific expert SCM.

Forthcoming work will pursue three key directions:

1. **Automated Causal Representation Learning**: Integrating differentiable causal structure discovery (e.g., NOTEARS-based methods) with CPO to infer SCMs from observational data, reducing manual specification burdens.
2. **Multi-Agent CPO**: Extending the framework to collaborative and competitive settings (e.g., autonomous fleets), where emergent behaviors demand counterfactual reasoning about other agents' intents.
3. **Resource-Aware Causal Compression**: Developing lightweight SCM approximations using knowledge distillation (Du et al., 2025) to deploy CPO on edge devices with <50ms latency.

Ultimately, CPO moves past conventional explainable RL by ensuring causal auditability to be a core property of policy optimization, no longer an external feature. This reduces independent systems from still opaque black boxes into accountable repositories whose choices can be scrutinized, understood, and relied upon. As regulations, like the EU AI Act (2024) mandate causal transparency for safety-critical AI, CPO provides the technical foundation for a whole new generation of ethically deployable autonomy.

References:

1. Abbott, S., Sherratt, K., Gerstung, M., & Funk, S. (2022). Estimation of the test-to-test distribution as a proxy for generation interval distribution for the Omicron variant in England. MedRxiv, 2022-01.
2. Appuhamilage, P. M. M. B. (2021). Explainable Reinforcement Learning Through a Causal Lens.
3. Assaad, M., Boné, R., & Cardot, H. (2008). A New Boosting Algorithm for Improved Time-Series Forecasting with Recurrent Neural Networks. *Information Fusion*, *9*(1). https://doi.org/10.1016/j.inffus.2006.10.009
4. Assaad, M. A., & Shakah, G. H. (2024). Optimizing Health Pattern Recognition Particle Swarm Optimization Approach for Enhanced Neural Network Performance. *Cihan University-Erbil Scientific Journal*, *8*(2), 76-83. https://doi.org/10.24086/cuesj.v8n2y2024.pp76-83
5. Assaad, M. A., Saleh, M. I., & Mahrousseh, R. (2025). A Novel Framework for Accurate Brain Tumor Detection in MRI Scans Using CNN, MLP, and KNN Techniques. *Journal of Information Hiding and Multimedia Signal Processing*, *16*(2).
6. Cheng, J., Chen, Y., Mei, X., Yang, B., Li, B., & Liu, M. (2024, May). Rethinking imitation-based planners for autonomous driving. In 2024 IEEE International Conference on Robotics and Automation (ICRA) (pp. 14123-14130). IEEE.

Received: August 07, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

7. Cheng, Z., Yu, J., & Xing, X. (2025). A Survey on Explainable Deep Reinforcement Learning. arXiv preprint arXiv:2502.06869.
8. Chen, L., Zhu, C., Soselia, D., Chen, J., Zhou, T., Goldstein, T., ... & Catanzaro, B. (2024). Odin: Disentangled reward mitigates hacking in rlhf. arXiv preprint arXiv:2402.07319.
9. Dazeley, R., Vamplew, P., & Cruz, F. (2023). Explainable reinforcement learning for broad-xai: a conceptual framework and survey. Neural Computing and Applications, 35(23), 16893-16916.
10. Deng, Z., Jiang, J., Long, G., & Zhang, C. (2023). Causal reinforcement learning: A survey. arXiv preprint arXiv:2307.01452.
11. Du, H., Wang, W., Zhang, W., Su, D., Bai, L., Bai, L., & Liang, J. (2025). Learning robust MLPs on graphs via cross-layer distillation from a causal perspective. Pattern Recognition, 111367.
12. EU AI Act. (2024). Regulation on Artificial Intelligence. Official Journal of the European Union.
13. Gu, Y., Cheng, Y., Chen, C. P., & Wang, X. (2021). Proximal policy optimization with policy feedback. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 52(7), 4600-4610.
14. Hu, K., Xu, K., Xia, Q., Li, M., Song, Z., Song, L., & Sun, N. (2024). An overview: Attention mechanisms in multi-agent reinforcement learning. Neurocomputing, 128015.
15. Hu, X., Zhang, R., Tang, K., Guo, J., Yi, Q., Chen, R., ... & Chen, Y. (2022). Causality-driven hierarchical structure discovery for reinforcement learning. Advances in neural information processing systems, 35, 20064-20076.
16. Hu, J., Stone, P., & Martín-Martín, R. (2023). Causal policy gradient for whole-body mobile manipulation. arXiv preprint arXiv:2305.04866.
17. Ji, J., Zhang, B., Zhou, J., Pan, X., Huang, W., Sun, R., ... & Yang, Y. (2023). Safety gymnasium: A unified safe reinforcement learning benchmark. Advances in Neural Information Processing Systems, 36, 18964-18993.
18. Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Lyu, Z., ... & Schölkopf, B. (2023). Cladder: Assessing causal reasoning in language models. Advances in Neural Information Processing Systems, 36, 31038-31065.
19. Kacianka, S., & Pretschner, A. (2021, March). Designing accountable systems. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 424-437).
20. Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A. A., Yogamani, S., & Pérez, P. (2021). Deep reinforcement learning for autonomous driving: A survey. IEEE transactions on intelligent transportation systems, 23(6), 4909-4926.
21. Lehmann, M. (2024). The definitive guide to policy gradients in deep reinforcement learning: Theory, algorithms and implementations. arXiv preprint arXiv:2401.13662.
22. Meng, W., Zheng, Q., Shi, Y., & Pan, G. (2021). An off-policy trust region policy optimization method with monotonic improvement guarantee for deep reinforcement learning. IEEE Transactions on Neural Networks and Learning Systems, 33(5), 2223-2235.
23. Min, C., Liao, G., Wen, G., Li, Y., & Guo, X. (2023). Ensemble Interpretation: A Unified Method for Interpretable Machine Learning. arXiv preprint arXiv:2312.06255.
24. Morales, E. F., Murrieta-Cid, R., Becerra, I., & Esquivel-Basaldua, M. A. (2021). A survey on deep learning and deep reinforcement learning in robotics with a tutorial on deep reinforcement learning. Intelligent Service Robotics, 14(5), 773-805.

Received: August 07, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

25. Polat Erdeniz, S., Veeranki, S., Schrempf, M., Jauk, S., Ngoc Trang Tran, T., Felfernig, A., ... & Leodolter, W. (2022, September). Explaining machine learning predictions of decision support systems in healthcare. In Current Directions in Biomedical Engineering (Vol. 8, No. 2, pp. 117-120). De Gruyter.
26. Rouff, C., & Watkins, L. (2022). Assured autonomy survey. Foundations and Trends® in Privacy and Security, 4(1), 1-116.
27. Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. Proceedings of the IEEE, 109(5), 612-634.
28. Sun, C., Zhang, R., Lu, Y., Cui, Y., Deng, Z., Cao, D., & Khajepour, A. (2023). Toward ensuring safety for autonomous driving perception: Standardization progress, research advances, and perspectives. IEEE Transactions on Intelligent Transportation Systems, 25(5), 3286-3304.
29. Vaskov, S., Schwarting, W., & Baker, C. (2024, June). Do no harm: A counterfactual approach to safe reinforcement learning. In 6th Annual Learning for Dynamics & Control Conference (pp. 1675-1687). PMLR.
30. Varadarajan, B., Hefny, A., Srivastava, A., Refaat, K. S., Nayakanti, N., Cornman, A., ... & Sapp, B. (2022, May). Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction. In 2022 International Conference on Robotics and Automation (ICRA) (pp. 7814-7821). IEEE.
31. Wang, C., & Aouf, N. (2024). Explainable deep adversarial reinforcement learning approach for robust autonomous driving. IEEE Transactions on Intelligent Vehicles.
32. Wu, J., Zhou, Y., Wang, H., Wang, X., & Wang, J. (2023). Assessing the causal effects of climate change on vegetation dynamics in Northeast China using convergence cross-mapping. IEEE Access, 11, 115367-115379.
33. Zeng, Y., Cai, R., Sun, F., Huang, L., & Hao, Z. (2024). A survey on causal reinforcement learning. IEEE Transactions on Neural Networks and Learning Systems.
34. Zhang, H., & Han, X. (2024). Off-policy asymptotic and adaptive maximum entropy deep reinforcement learning. International Journal of Machine Learning and Cybernetics, 1-13.

Received: August 07, 2025