Novelty Assessment Report

Paper: How to Lose Inherent Counterfactuality in Reinforcement Learning
PDF URL: https://openreview.net/pdf?id=2kutK2Y8Sv
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-05

Abstract

Learning in high-dimensional MDPs with complex state dynamics became possible with the progress achieved in reinforcement learning research. At the same time, deep neural policies have been observed to be highly unstable with respect to the minor variations in their state space, causing volatile and unpredictable behaviour. To alleviate these volatilities, a line of work suggested techniques to cope with this problem via explicitly regularizing the temporal difference loss to ensure local ϵ-invariance in the state space. In this paper, we provide theoretical foundations on the impact of ϵ-local invariance training on the deep neural policy manifolds. Our comprehensive theoretical and experimental analysis reveals that standard reinforcement learning inherently learns counterfactual values while recent training techniques that focus on explicitly enforcing ϵ-local invariance cause policies to lose counterfactuality, and further result in learning misaligned and inconsistent values. In connection to this analysis, we further highlight that this line of training methods break the core intuition and the true biological inspiration of reinforcement learning, and introduce an intrinsic gap between how natural intelligence understands and interacts with an environment in contrast to AI agents trained via ϵ-local invariance methods. The misalignment, inaccuracy and the loss of counterfactuality revealed in our paper further demonstrate the need to rethink the approach in establishing truly reliable and generalizable reinforcement learning policies.

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: **Impact of Epsilon-Local Invariance Training on Reinforcement Learning Policies**

A total of **1 papers** were analyzed and organized into a taxonomy with **2 categories**.

Taxonomy Overview

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

- **Theoretical Foundations and Counterfactual Analysis**
- **Training Methods and Robustness**

Complete Taxonomy Tree

- Impact of Epsilon-Local Invariance Training on Reinforcement Learning Policies Survey Taxonomy
- Theoretical Foundations and Counterfactual Analysis
  - Counterfactual Value Learning and Policy Manifolds ★ (1 papers)
  - [0] How to Lose Inherent Counterfactuality in Reinforcement Learning (Anon et al., 2026) View paper
- Training Methods and Robustness
  - Adversarial Robustness Training (1 papers)
  - [1] An active learning framework for adversarial training of deep neural networks (Susmita Ghosh, 2025) View paper

Narrative

Core task: impact of epsilon-local invariance training on reinforcement learning policies. The field structure suggested by this taxonomy divides into two main branches. The first, Theoretical Foundations and Counterfactual Analysis, examines the conceptual underpinnings of how policies represent and leverage counterfactual reasoning—essentially, understanding what alternative actions might have yielded and how epsilon-local constraints shape the manifold of learned policies. The second branch, Training Methods and Robustness, focuses on practical algorithms and techniques for building policies that remain stable under small perturbations, often through adversarial or regularization-based approaches. Together, these branches capture both the 'why' and the 'how' of epsilon-local invariance: one side investigates the theoretical implications for value learning and policy geometry, while the other develops concrete training recipes to achieve robustness.

A particularly active line of work explores the tension between enforcing local invariance and preserving the policy's ability to distinguish meaningful state differences. Losing Inherent Counterfactuality[0] sits within the Counterfactual Value Learning and Policy Manifolds cluster, emphasizing how epsilon-local training can inadvertently suppress the counterfactual signals that guide effective exploration and credit assignment. This contrasts with approaches like Active Adversarial Training[1], which prioritize robustness by explicitly injecting perturbations during learning, potentially at the cost of nuanced counterfactual reasoning. The central open question is whether one can design training schemes that simultaneously maintain local invariance for robustness and retain the rich counterfactual structure needed for sample-efficient learning, or whether these goals inherently trade off against one another.

Related Works in Same Category

No sibling papers and no sibling subtopics were found under the same parent taxonomy node; the paper appears structurally isolated in the taxonomy.

Contributions Analysis

Overall novelty summary. The paper investigates how epsilon-local invariance training affects deep reinforcement learning policies, focusing on counterfactual value learning and policy manifold geometry. Within the taxonomy, it occupies the 'Counterfactual Value Learning and Policy Manifolds' leaf under 'Theoretical Foundations and Counterfactual Analysis'. Notably, this leaf contains only the

original paper itself, with no sibling papers identified. This positioning suggests the work addresses a relatively sparse research direction, examining theoretical properties that have received limited direct attention in the literature surveyed.

The taxonomy reveals a clear structural division: theoretical foundations examining counterfactual reasoning versus practical training methods emphasizing adversarial robustness. The original paper sits in the former branch, while the neighboring 'Adversarial Robustness Training' leaf (containing one paper on active adversarial training) represents the practical counterpart. The taxonomy's scope notes explicitly separate theoretical counterfactual analysis from adversarial training techniques, indicating these represent distinct but complementary research threads. The paper's focus on manifold geometry and value alignment positions it at the conceptual foundation of understanding epsilon-local constraints, rather than in the algorithmic development space.

Among twenty candidates examined across three contributions, none were identified as clearly refuting the paper's claims. The second contribution (inherent counterfactual reasoning in standard RL) examined ten candidates with zero refutable matches, as did the third contribution (counterfactuality-robustness trade-off). The first contribution (theoretical analysis of epsilon-local invariance effects) examined zero candidates. This pattern suggests that within the limited search scope of top-K semantic matches, the specific theoretical framing around counterfactual loss and policy manifolds appears relatively unexplored. However, the small candidate pool (twenty papers total) means the analysis covers a narrow slice of potentially relevant work.

Given the limited search scope and sparse taxonomy structure, the work appears to occupy a relatively novel theoretical niche within the examined literature. The absence of sibling papers and refutable candidates among twenty examined suggests the specific angle—connecting epsilon-local invariance to counterfactual value learning—has not been directly addressed in closely related work. However, this assessment is constrained by the top-K semantic search methodology and does not reflect an exhaustive survey of reinforcement learning robustness or theoretical RL literature more broadly.

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

Contribution 1: Theoretical analysis of ε-local invariance training effects on Q-functions

Description: The authors present a formal theoretical framework demonstrating that ε-local invariance training fundamentally alters learned value judgments in reinforcement learning. They prove an inherent trade-off between accurate Q-value estimation and robustness, showing that ε-invariant Q-functions overestimate optimal values and misalign counterfactual action rankings.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Contribution 2: Discovery that standard RL possesses inherent counterfactual reasoning ability

Description: The authors establish that standard reinforcement learning naturally learns counterfactual values aligned with human decision-making processes, while ε-invariance training methods cause policies to lose this inherent counterfactual ability, resulting in inaccurate, inconsistent, and misaligned value functions.

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. Fast Counterfactual Inference for History-Based Reinforcement Learning

URL: View paper

Brief Assessment

Fast Counterfactual Inference[6] focuses on counterfactual inference as a computational technique for history compression in partially-observable tasks, not on analyzing whether standard RL inherently learns counterfactual values or comparing this ability against ε-invariance training methods.

---

2. Do No Harm: A Counterfactual Approach to Safe Reinforcement Learning

URL: View paper

Brief Assessment

Do No Harm[2] focuses on formulating counterfactual harm constraints for safe RL by comparing learned policies to default safe policies, not on analyzing whether standard RL inherently learns counterfactual values or how training methods affect this ability.

---

3. Reward Machines: Exploiting Reward Function Structure in Reinforcement Learning

URL: View paper

Brief Assessment

Reward Machines[3] focuses on exploiting reward function structure through finite state machines for sample efficiency, not on analyzing counterfactual reasoning abilities inherent to standard RL methods or comparing them with ε-invariance training approaches.

---

4. Shapley Counterfactual Credits for Multi-Agent Reinforcement Learning

URL: View paper

Brief Assessment

Shapley Counterfactual Credits[8] focuses on credit assignment in multi-agent RL using Shapley values to distribute rewards among cooperating agents, not on analyzing whether standard RL inherently learns counterfactual values or comparing standard RL to epsilon-invariance training methods.

---

5. Reasoning about Counterfactuals to Improve Human Inverse Reinforcement Learning

URL: View paper

Brief Assessment

Counterfactuals Human IRL[5] focuses on human inverse reinforcement learning and how robots can select informative demonstrations by reasoning about human counterfactuals. It does not address the inherent counterfactual reasoning ability of standard RL algorithms themselves, which is the core claim of the original paper.

---

6. On Minimizing Adversarial Counterfactual Error in Adversarial Reinforcement Learning

URL: View paper

Brief Assessment

Minimizing Adversarial Counterfactual[9] focuses on adversarial counterfactual error in partially observable settings under adversarial perturbations, not on analyzing inherent counterfactual reasoning abilities of standard RL.

---

7. Counterfactual influence in Markov decision processes

URL: View paper

Brief Assessment

Counterfactual Influence MDPs[11] focuses on counterfactual inference in MDPs using structural causal models (SCMs) to derive what-if scenarios under different action sequences. The original paper claims standard RL inherently learns counterfactual values aligned with human decision-making, while Counterfactual Influence MDPs[11] addresses a different problem: ensuring that counterfactual paths remain influenced by observed paths when using Gumbel-max SCMs for counterfactual inference. These are distinct research directions within counterfactual reasoning.

---

8. Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning

URL: View paper

Brief Assessment

Social Influence Motivation[4] focuses on multi-agent coordination through causal influence rewards and counterfactual reasoning about other agents' actions, not on analyzing standard RL's inherent counterfactual abilities in single-agent settings or comparing it against ε-invariance training methods.

---

9. SAFE-RL: Saliency-Aware Counterfactual Explainer for Deep Reinforcement Learning Policies

URL: View paper

Brief Assessment

SAFE-RL[7] focuses on generating counterfactual explanations for trained DRL policies using saliency-aware GANs, not on analyzing whether standard RL inherently learns counterfactual values during training. The paper addresses explainability of existing policies rather than the learning dynamics of RL algorithms.

---

10. Eliciting Chain-of-Thought Reasoning for Time Series Analysis using Reinforcement Learning

URL: View paper

Brief Assessment

Chain-of-Thought Time Series[10] focuses on training LLMs for time series analysis using RL with chain-of-thought reasoning. It does not address counterfactual reasoning abilities inherent to standard RL algorithms or compare standard RL with ε-invariance training methods.

Contribution 3: Identification of fundamental trade-off between counterfactuality and robustness

Description: The authors formalize and demonstrate through theory and experiments a fundamental trade-off showing that certified ε-invariance training sacrifices the inherent counterfactual reasoning capabilities of standard RL in pursuit of robustness guarantees, revealing core mechanisms behind this phenomenon.

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. User Retention-oriented Recommendation with Decision Transformer

URL: View paper

Brief Assessment

Retention Decision Transformer[20] focuses on user retention optimization in recommendation systems using decision transformers, not on the trade-off between counterfactuality and robustness in reinforcement learning training methods.

---

2. Budgeting Counterfactual for Offline RL

URL: View paper

Brief Assessment

Budgeting Counterfactual Offline[16] focuses on offline RL with limited data and counterfactual reasoning about alternative actions, not on the trade-off between certified ε-invariance training and counterfactual capabilities in online RL settings.

---

3. Regret Minimization for Partially Observable Deep Reinforcement Learning

URL: View paper

Brief Assessment

Regret Minimization POMDP[15] focuses on counterfactual regret minimization for handling partial observability in POMDPs, not on analyzing trade-offs between counterfactuality and robustness guarantees in adversarial training contexts. The candidate's counterfactual framework addresses non-Markovian observations, while the original examines how ε-invariance training affects counterfactual reasoning capabilities.

---

4. Causal Counterfactuals for Improving the Robustness of Reinforcement Learning

URL: View paper

Brief Assessment

Causal Counterfactuals Robustness[14] focuses on using causal counterfactuals to improve robustness in robotic manipulation tasks, not on analyzing trade-offs between counterfactuality and certified ε-invariance training methods in RL.

---

5. On Minimizing Adversarial Counterfactual Error in Adversarial Reinforcement Learning

URL: View paper

Brief Assessment

Minimizing Adversarial Counterfactual[9] proposes a new objective (ACOE) to balance value optimization with robustness, rather than identifying or formalizing a fundamental trade-off between counterfactuality and robustness in ε-invariance training.

---

6. Promoting counterfactual robustness through diversity

URL: View paper

Brief Assessment

Counterfactual Robustness Diversity[12] addresses robustness of counterfactual explanations in explainable AI (XAI), not reinforcement learning. The candidate focuses on how minor input changes affect explanation stability in classification models, while the original paper examines trade-offs between counterfactual reasoning and ε-invariance training in RL policies.

---

7. Generating robust counterfactual explanations

URL: View paper

Brief Assessment

Robust Counterfactual Explanations[19] addresses robustness in counterfactual explanations for classification models, not reinforcement learning policies. The paper focuses on trade-offs between proximity and robustness in generating counterfactual explanations, which is a different domain from RL's counterfactual reasoning capabilities.

---

8. Robust Counterfactual Inference in Markov Decision Processes

URL: View paper

Brief Assessment

Robust Counterfactual MDPs[13] addresses uncertainty in causal models for counterfactual inference and robust policy optimization, not the trade-off between counterfactual reasoning capabilities and robustness guarantees in RL training methods.

---

9. Bayesian Uncertainty Estimation for Targeted Counterfactual Experience Generation in Reinforcement Learning

URL: View paper

Brief Assessment

Bayesian Counterfactual Generation[18] focuses on counterfactual experience generation for data utilization in off-policy RL, not on the trade-off between counterfactuality and robustness guarantees in certified ε-invariance training.

---

10. Masked Images Are Counterfactual Samples for Robust Fine-Tuning

URL: View paper

Brief Assessment

Masked Counterfactual Samples[17] addresses a trade-off between in-distribution performance and out-of-distribution robustness in fine-tuning vision models, not the fundamental trade-off between counterfactual reasoning capabilities and certified ε-invariance training in reinforcement learning policies.

Appendix: Text Similarity Detection

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

References

- [0] How to Lose Inherent Counterfactuality in Reinforcement Learning View paper
- [1] An active learning framework for adversarial training of deep neural networks View paper
- [2] Do No Harm: A Counterfactual Approach to Safe Reinforcement Learning View paper
- [3] Reward Machines: Exploiting Reward Function Structure in Reinforcement Learning View paper
- [4] Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning View paper
- [5] Reasoning about Counterfactuals to Improve Human Inverse Reinforcement Learning View paper
- [6] Fast Counterfactual Inference for History-Based Reinforcement Learning View paper
- [7] SAFE-RL: Saliency-Aware Counterfactual Explainer for Deep Reinforcement Learning Policies View paper
- [8] Shapley Counterfactual Credits for Multi-Agent Reinforcement Learning View paper
- [9] On Minimizing Adversarial Counterfactual Error in Adversarial Reinforcement Learning View paper
- [10] Eliciting Chain-of-Thought Reasoning for Time Series Analysis using Reinforcement Learning View paper
- [11] Counterfactual influence in Markov decision processes View paper
- [12] Promoting counterfactual robustness through diversity View paper
- [13] Robust Counterfactual Inference in Markov Decision Processes View paper
- [14] Causal Counterfactuals for Improving the Robustness of Reinforcement Learning View paper
- [15] Regret Minimization for Partially Observable Deep Reinforcement Learning View paper
- [16] Budgeting Counterfactual for Offline RL View paper
- [17] Masked Images Are Counterfactual Samples for Robust Fine-Tuning View paper
- [18] Bayesian Uncertainty Estimation for Targeted Counterfactual Experience Generation in Reinforcement Learning View paper
- [19] Generating robust counterfactual explanations View paper
- [20] User Retention-oriented Recommendation with Decision Transformer View paper