Reinforcement Learning Security and Safety

Ezgi Korkmaz

February 1, 2026

[1]Ezgi Korkmaz. Counteractive RL: Rethinking Core Principles for Efficient and Scalable Deep Reinforcement Learning. Conference on Neural Information Processing Systems (NeurIPS), **NeurIPS 2025**. **Spotlight Presentation**

[2]Ezgi Korkmaz. How to Lose Inherent Counterfactuality in Reinforcement Learning. International Conference on Learning Representations, (ICLR), **ICLR 2026**.

[3]Ezgi Korkmaz. Fair Reinforcement Learning. International Conference on Learning Representations, (ICLR), **ICLR 2026**.

[4] Ezgi Korkmaz. Principled Analysis of Deep Reinforcement Learning Design and Evaluation Paradigms. AAAI Conference on Artificial Intelligence **[Acceptance Rate: 17.6%]**, **AAAI 2026**.

[5] Ezgi Korkmaz. Understanding and Diagnosing Deep Reinforcement Learning. International Conference on Machine Learning **[Acceptance Rate: 27.54%]**, **ICML 2024**.

[6] Ezgi Korkmaz. Adversarial Robust Deep Reinforcement Learning Requires Redefining Robustness. AAAI Conference on Artificial Intelligence **[Acceptance Rate: 19.6%]**, **AAAI 2023**.

[7] Ezgi Korkmaz et al. Detecting Adversarial Directions in Deep Reinforcement Learning to Make Robust Decisions. International Conference on Machine Learning, **[Acceptance Rate: 27.94%]**, **ICML 2023**.

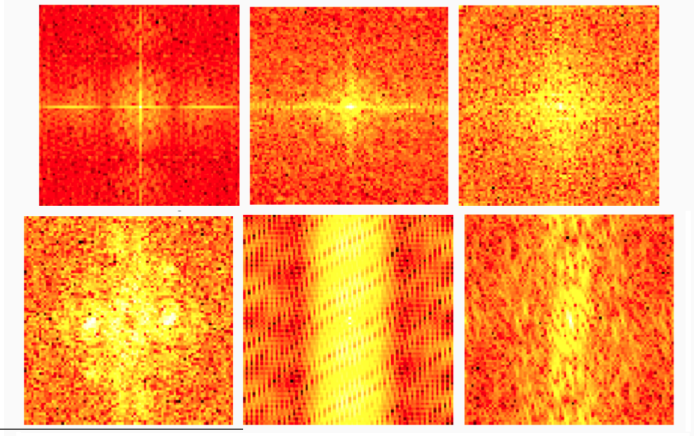
[8] Ezgi Korkmaz. Deep Reinforcement Learning Policies Learn Shared Adversarial Features Across MDPs. AAAI Conference on Artificial Intelligence **[Acceptance Rate: 14.58%]**, **AAAI 2022**.

[9] Ezgi Korkmaz. Investigating Vulnerabilities of Deep Neural Policies. Conference on Uncertainty in Artificial Intelligence (UAI) **[Acceptance Rate: 26.4%]**, **UAI 2021**.

[10] Ezgi Korkmaz. Machine Learning Safety: From Reinforcement Learning to Foundation Models. AAAI Conference on Artificial Intelligence Tutorial, **AAAI 2025**.

How much we can trust the decisions?

This study [10] reveals that robust reinforcement learning is not robust.



[9] Ezgi Korkmaz. Investigating Vulnerabilities of Deep Neural Policies. Conference on Uncertainty in Artificial Intelligence (UAI) **[Acceptance Rate: 26.4%]**, **UAI 2021**

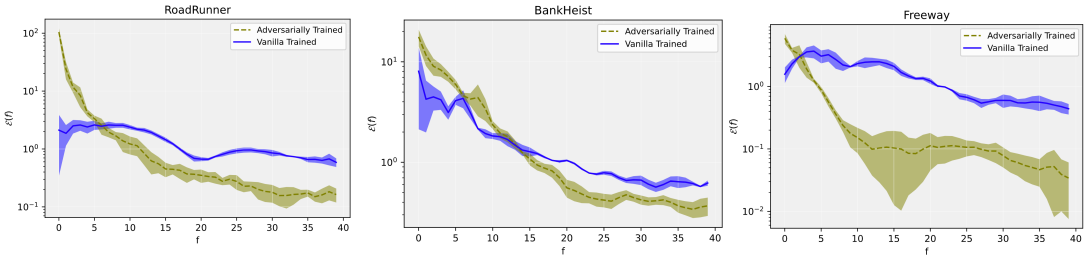
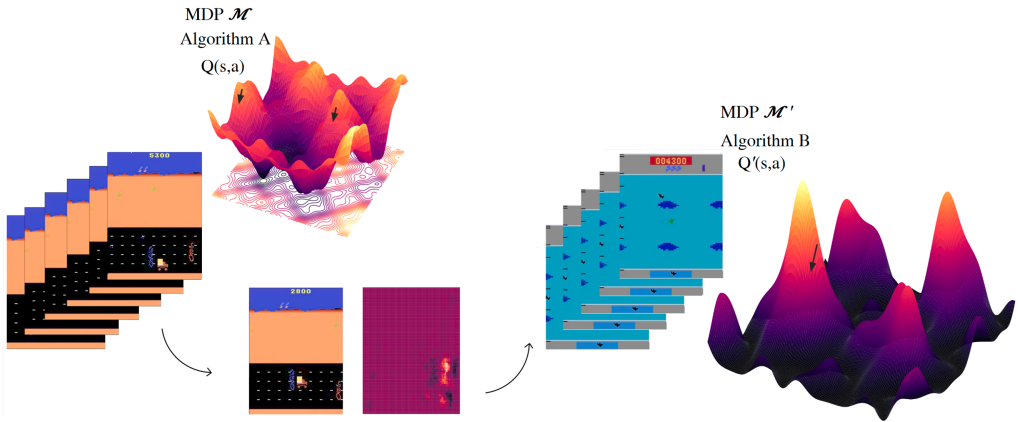


Figure 1: Power spectrum $\mathcal{E}(f)$ of the perturbations computed via the Carlini & Wagner formulation for the certified robust trained models and vanilla trained models in the Fourier domain.

[9] Ezgi Korkmaz. Investigating Vulnerabilities of Deep Neural Policies. Conference on Uncertainty in Artificial Intelligence (UAI) **[Acceptance Rate: 26.4%]**, **UAI 2021**

Transferability and Inevitability

AAAI 2022 *paper [8] shows that adversarial perturbations transfers across states, across MDPs and across algorithms and introduces the theoretical foundations that shows inevitability of high-sensitivity directions in high dimensional MDPs.*



[8] Ezgi Korkmaz. Deep Reinforcement Learning Policies Learn Shared Adversarial Features Across MDPs. AAAI Conference on Artificial Intelligence, **AAAI 2022**.

AI Security and Safety Implications

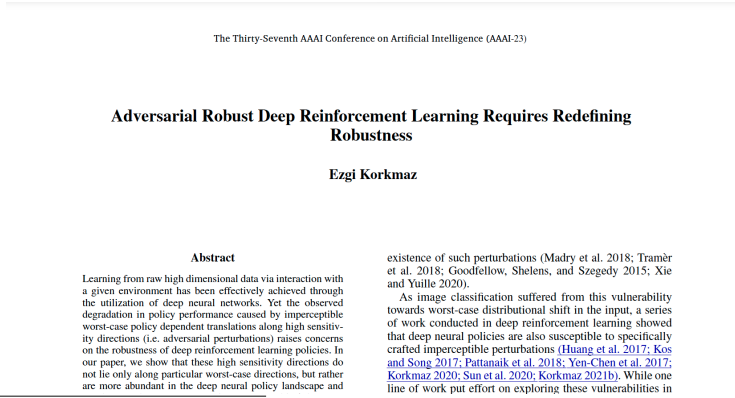
The adversary does **not need to know**!

- Training details
- Training algorithm
- Training network
- Training MDP

A complete black-box attack

Generalization and Robustness

AAAI 2023 *paper [6] reveals that robust reinforcement learning does not generalize even as much as standard reinforcement learning can and the extensive line of research in current machine learning does not even ask the right research question.*



The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)

Adversarial Robust Deep Reinforcement Learning Requires Redefining Robustness

Ezgi Korkmaz

Abstract

Learning from raw high dimensional data via interaction with a given environment has been effectively achieved through the utilization of deep neural networks. Yet the observed degradation in policy performance caused by imperceptible worst-case policy dependent translations along high sensitivity directions (i.e. adversarial perturbations) raises concerns on the robustness of deep reinforcement learning policies. In our paper, we show that these high sensitivity directions do not lie only along particular worst-case directions, but rather are more abundant in the deep neural policy landscape and

existence of such perturbations (Madry et al. 2018; Tramèr et al. 2018; Goodfellow, Shelens, and Szegedy 2015; Xie and Yuille 2020).

As image classification suffered from this vulnerability towards worst-case distributional shift in the input, a series of work conducted in deep reinforcement learning showed that deep neural policies are also susceptible to specifically crafted imperceptible perturbations (Huang et al. 2017; Kos and Song 2017; Pattanaik et al. 2018; Yen-Chen et al. 2017; Korkmaz 2020; Sun et al. 2020; Korkmaz 2021b). While one line of work put effort on exploring these vulnerabilities in

[6] Ezgi Korkmaz. Adversarial Robust Deep Reinforcement Learning Requires Redefining Robustness. AAAI Conference on Artificial Intelligence **AAAI 2023**.

AI Security and Safety Implications

The adversary does **not need to know**!

- Not only the training policy
- But any policy
- Attacks without accessing policies
- Without computing gradients

Security via Analyzing Deep Neural Policy Manifold

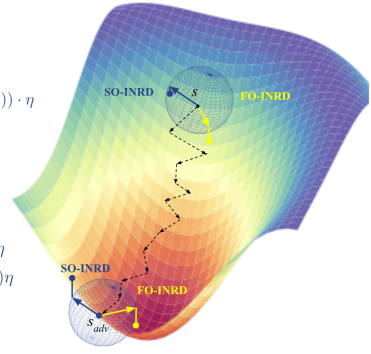
*My **ICML 2023** paper [7] introduces foundational methods to identify and detect the limitations of deep reinforcement learning by leveraging the structure of deep neural manifolds.*

$$\tilde{J}(s_0, \eta) = J(s_0, \pi^*(\cdot \mid s_0)) + \nabla_s J(s_0, \pi^*(\cdot \mid s_0)) \cdot \eta$$

$$\mathcal{L}(s_0, \eta) = J(s_0 + \eta, \pi^*(\cdot \mid s_0)) - \tilde{J}(s_0, \eta)$$

$$\begin{aligned} J(s_0 + \eta, \pi^*(\cdot|s_0)) \approx & J(s_0, \pi^*(\cdot|s_0)) \\ & + \nabla_s J(s_0, \pi^*(\cdot|s_0)) \cdot \eta \\ & + \eta^\top \nabla_s^2 J(s_0, \pi^*(\cdot|s_0))\eta \end{aligned}$$

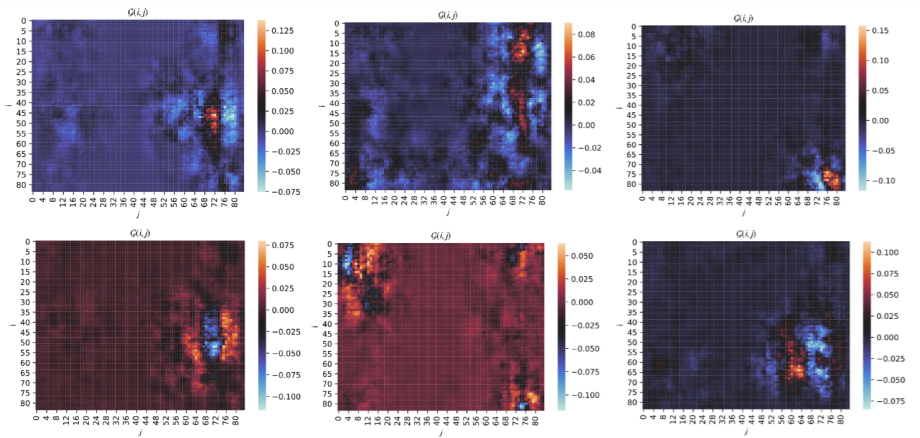
$$\begin{aligned} \mathcal{L}(s_0, \eta) &\approx \eta^\top \nabla_s^2 J(s_0, \pi^*(\cdot|s_0))\eta \\ &\geq \lambda_{\min}\left(\nabla_s^2 J(s, \tau)\right) \|\eta\|^2 \end{aligned}$$



$$\mathcal{K}(s_0, \eta) = J(s_0 + \eta, \pi^*(\cdot \mid s_0)) - J(s_0, \pi^*(\cdot \mid s_0))$$

[7] Ezgi Korkmaz et al. Detecting Adversarial Directions in Deep Reinforcement Learning to Make Robust Decisions. International Conference on Machine Learning, **[Acceptance Rate: 27.94%]**, **ICML 2023**.

Deep Reinforcement Learning Decision Making



[5] Ezgi Korkmaz. Understanding and Diagnosing Deep Reinforcement Learning. International Conference on Machine Learning **ICML 2024**.