Risk-calibrated Robust Adversarial Reinforcement Learning

Ziyao Mou zmou1@jh.edu

Shuwen Bai sbai12@jh.edu

December 9, 2025

1. Introduction

Robust Reinforcement Learning seeks to improve a policy's performance under worst-case perturbations. Among these approaches, Robust Adversarial Reinforcement Learning (RARL)[1] has emerged as a widely adopted framework, where an explicitly trained adversary generates harmful perturbations while the protagonist learns to withstand them. Despite its conceptual simplicity, RARL exhibits two fundamental issues. First, the adversary may behave noisily: it frequently attacks in low-risk states where disturbances are unnecessary and provide little training value. Second, due to sparse failure signals, the adversary may fail to focus its attack budget on truly high-risk states where perturbations matter most. These behaviors reduce both robustness and sample efficiency, and in extreme cases may even destabilize training.

To address these limitations, we introduce Risk-calibrated Robust Adversarial Reinforcement Learning (RC-RARL), a simple yet effective modification of the RARL framework. Our key idea is to learn a probabilistic risk model that predicts the likelihood of failure, and then apply post-hoc calibration techniques to correct systematic miscalibration in these probability estimates. A calibrated risk model provides a reliable estimate of the true danger level of each state, enabling the adversary to selectively apply perturbations only when the agent is genuinely close to failing. This results in a more targeted and efficient adversarial signal, reducing unnecessary disturbances while strengthening robustness in high-risk regions.

We evaluate the proposed method on standard continuous-control environments and observe two consistent benefits. First, the calibrated risk model yields more reliable probability estimates, reflected in lower ECE and smoother reliability curves. Second, the risk-calibrated adversary leads to higher final rewards and faster convergence compared to standard RARL, indicating more focused and meaningful adversarial interactions. These results highlight the importance of calibration in adversarial RL and point toward a promising direction for developing more principled, risk-aware robust RL algorithms.

2. Related Work

2.1. Adversarial Training for Robust RL

Robust Adversarial Reinforcement Learning (RARL) introduces an explicitly trained adversary that learns to generate harmful perturbations against the protagonist policy[1]. This min–max formulation creates worst-case training conditions that improve policy resilience and test-time robustness. Several extensions have improved adversary design, such as learning stronger perturbation distributions, state-dependent adversaries, or distributionally robust adversarial agents. Despite these advances, existing adversarial RRL methods primarily focus on maximizing worst-case returns and do not model uncertainty or reliability in the learned perturbation distributions.

2.2. Uncertainty Estimation and Calibration

Accurate uncertainty estimation is crucial for decision-making. Plenty of work studies miscalibration in modern neural networks[2] and develops methods such as Platt scaling, temperature scaling, and isotonic regression to improve probabilistic reliability. In RL, uncertainty has been used for exploration, model-based planning, and robust control, but calibration has received limited attention. Malik et al.[3] demonstrated that calibrated uncertainty significantly improves model-based RL performance, enabling better planning and reducing model-bias errors. Other recent work shows that adversarial robustness and uncertainty calibration are closely connected, and adversarial training can implicitly affect calibration quality[4].

3. Proposed Approach

We introduce Risk-Calibrated Robust Adversarial Reinforcement Learning (RC-RARL). Our method preserves the min–max training structure of RARL but modifies how the adversary decides when to attack, enabling more targeted perturbations and reducing unnecessary disturbances.

3.1. Adversarial Training in RARL

RARL formulates robustness as a two-player zero-sum game between a protagonist policy π_P that maximizes return and an adversary policy π_A that minimizes it. Training proceeds through alternating optimization, where π_P is updated using TRPO while holding π_A fixed, and then π_A is updated to generate worst-case disturbances under a fixed protagonist. Although this framework improves robustness, it often leads to noisy adversarial behavior: the adversary frequently attacks in low-risk states and fails to concentrate its disturbance budget on genuinely dangerous states. This issue arises primarily from the sparsity of failure signals, which provide weak supervision and prevent the adversary from learning when perturbations are truly necessary.

3.2. Learning a Failure Risk Model

To address this issue, we introduce a learned failure-risk estimator

$$r_\theta(s) \in [0, 1],$$

where $r_\theta(s)$ predicts the probability that a state s will lead to a failure within a short horizon.

Using trajectory rollouts, each visited state is assigned a binary label indicating whether a failure event occurred. We determine failure events by comparing the adversarial test returns against a dynamically estimated failure threshold, as detailed in Algorithm 1.

Algorithm 1 Adversary Failure Labeling and Data Collection

Input: Online training loop from RARL, adv_risk_update_every R, buffer D for states and labels, window size M for percentile estimation.

Initialize failure threshold $\tau_{\text{fail}} \leftarrow$ None

for *each global iteration* $k = 0, 1, \ldots, N_{itr}$ **do**

  Obtain adversarial test returns $\{\text{adv\_test}[i]\}_{i<k}$

  if τ_{fail} *is None and* k *has at least* M *test points* **then**

    Set failure threshold as the 20-th percentile of recent tests:

$$\tau_{\text{fail}} \leftarrow \text{Percentile}_\omega(\text{adv\_test}[k - M + 1 : k])$$

  end

  From protagonist TRPO, obtain trajectories $\{\tau_j = (s_t^{(j)}, a_t^{(j)}, r_t^{(j)})_{t=0}^{T_j-1}\}$, and label y_j to every s_j

end

We then fit a logistic regression model:

$$r_\theta(s) = \sigma(w^\top \phi(s)),$$

where $\phi(s)$ denotes state features. This provides a dense and informative failure signal that the adversary can leverage.

3.3. Calibration of the Risk Model

Raw probability outputs from logistic regression are often miscalibrated. To correct this, we apply post-hoc calibration via histogram binning. The calibration procedure is detailed in Algorithm 2.

Algorithm 2 Logistic Risk Modeling and Histogram Calibration

Input: Buffer D with state-label pairs, adv_risk_update_every R.

Initialize logistic model parameters (w, b) and histogram calibrator C

for *each global iteration* $k = 0, 1, \ldots, N_{itr}$ **do**

  if $k+1$ *is divisible by* R *and* D *is non-empty* **then**

    Stack all states and labels in D: $X \in \mathbb{R}^{N \times d}, y \in \{0,1\}^N$

    Fit logistic risk model:

$$p_{\text{raw}}(s) = \sigma(w^\top s + b),$$

    by minimizing cross-entropy via SGD on (X, y)

    Compute $p_{\text{raw}} = p_{\text{raw}}(X)$ on all collected states

    Fit histogram calibrator C:

    Partition $[0,1]$ into bins $\{B_k\}_{k=1}^K$

    For each bin B_k, set

$$\hat{p}_k = \mathbb{E}[y \mid p_{\text{raw}}(s) \in B_k]$$

    Define calibrated risk as

$$p_{\text{cal}}(s) = \mathcal{C}(p_{\text{raw}}(s)) = \hat{p}_k \textbf{ if } p_{\text{raw}}(s) \in B_k$$

  end

end

The calibrated risk is computed as:

$$\hat{r}(s) = \text{Calibrate}(r_\theta(s)),$$

where prediction values are grouped into bins and replaced with empirical failure frequencies.

This improves reliability (lower ECE, better calibration curves), producing more trustworthy risk estimates for controlling the adversary.

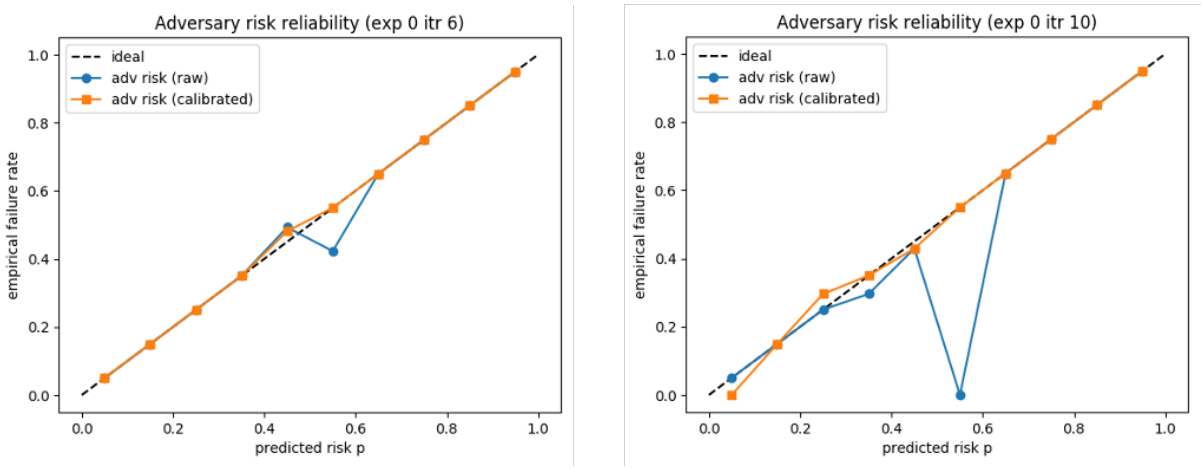
3.4. Risk-Aware Adversary Gating

The calibrated risk estimate $\hat{r}(s)$ is used to gate the adversary's action:

$$a_A(s) = \begin{cases} \pi_A(s), & \text{if } \hat{r}(s) > \tau, \\ 0, & \text{otherwise,} \end{cases}$$

where τ is a predefined risk threshold.

In this way, the adversary applies disturbances only when the agent is in a genuinely high-risk state, leading to more focused perturbations in dangerous regions, fewer unnecessary disturbances in safe regions, and a cleaner learning signal for both policies



(a) Iteration 6

(b) Iteration 10

Figure 1: Adversary risk reliability curves at different training stages. The curves show that calibration consistently aligns predicted risk with empirical failure frequency. As training progresses from iteration 6 to iteration 10, the calibrated model maintains a shape closer to the diagonal, indicating stable and improved probability reliability.

throughout training.

3.5. Training Loop

The full RC-RARL training loop is summarized as follows:

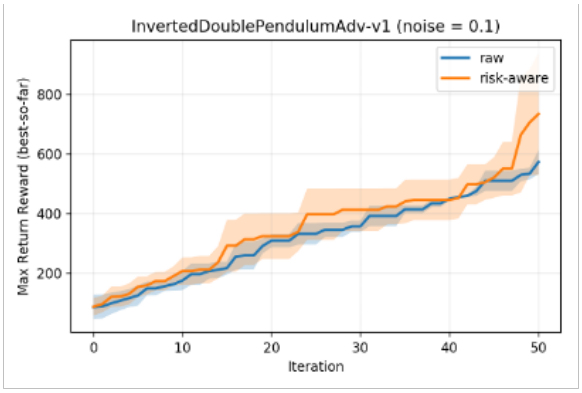
1. Run rollouts with the protagonist and gated adversary.
2. Update the risk model using collected state–failure pairs.
3. Apply histogram binning to calibrate risk predictions.
4. Update π_P using TRPO (fixing π_A).
5. Update π_A using TRPO with risk-aware gating.

This procedure integrates seamlessly into the original RARL framework while substantially improving the quality and relevance of adversarial disturbances.

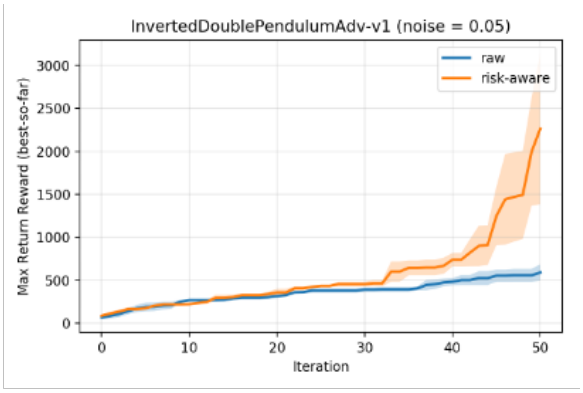
4. Results

4.1. Calibration Performence

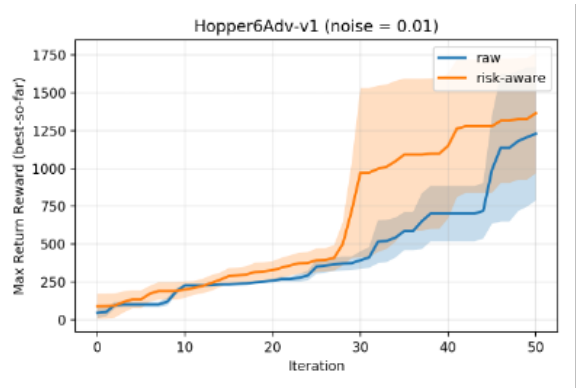
To assess how well the adversary's risk model reflects true failure likelihood, we visualize reliability curves at different stages of training (Figure 1). At iteration 6, the uncalibrated model already shows noticeable misalignment between predicted and empirical failure rates, whereas the calibrated model stays closer to the diagonal. By iteration 10, this pattern persists: calibration consistently corrects overconfidence and underconfidence in the raw risk estimates. This demonstrates that calibration improves probability reliability not only at convergence but also throughout training.



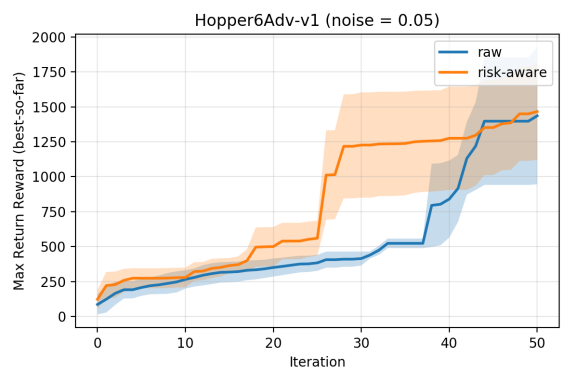
(a) InvertedDoublePendulumAdv-v1 (noise = 0.10)



(b) InvertedDoublePendulumAdv-v1 (noise = 0.05)



(c) Hopper6Adv-v1 (noise = 0.01)



(d) Hopper6Adv-v1 (noise = 0.05)

Figure 2: Max-return reward curves across training iterations for four environments and noise settings. The risk-calibrated adversary consistently achieves higher return and faster convergence compared to standard RARL, demonstrating improved robustness across different disturbance scales and task dynamics.

4.2. Reward Performance

Figure 2 presents the max-return reward curves across four environment–noise settings. Across all tasks, the risk-calibrated adversary achieves noticeably higher rewards compared to standard RARL, and the performance gap becomes more pronounced as training progresses. Notably, the improvement is more evident under smaller disturbance scales, suggesting that calibration is particularly effective when the environmental dynamics are less dominated by noise.

For InvertedDoublePendulumAdv-v1, calibration consistently enhances learning stability, with the advantage being especially clear at lower noise levels (noise = 0.05). For Hopper6Adv-v1, the calibrated adversary accelerates early-stage learning while also achieving higher final rewards in both noise regimes.

These results suggest that calibration helps the adversary focus its attack budget on high-risk states, strengthening the protagonist's robustness across diverse control tasks and disturbance intensities.

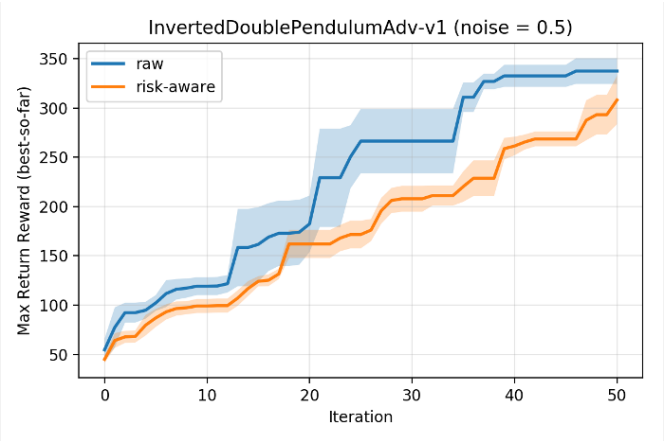


Figure 3: Reward comparison on InvertedDoublePendulumAdv-v1 with extremely high reset noise (0.50). In this setting, raw RARL outperforms RC-RARL throughout training.

5. discussion

While our results demonstrate clear benefits of risk calibration in several environments and training regimes, our method also presents important limitations. One limitation appears in the Hopper6Adv-v1 experiments, where RC-RARL achieves a noticeable advantage during iterations 20–40 but gradually converges to similar performance as the raw RARL baseline after iteration 40. This suggests that once the protagonist becomes sufficiently robust, the marginal influence of a calibrated adversary diminishes. As failure events decrease and trajectories become more stable, the risk model receives fewer informative samples, causing its guidance effect to weaken over time. Consequently, both the calibrated and uncalibrated adversaries may converge to similar long-term dynamics.

A second limitation emerges under extremely high disturbance levels, such as Inverted Double PendulumAdv-v1 with noise = 0.50, as illustrated in Figure 3. In this situation, the raw RARL baseline consistently outperforms our calibrated approach. When reset noise becomes excessively large, failures are dominated by environmental randomness rather than by meaningful state-dependent risk. As a result, the failure labels become highly noisy, preventing the risk model from learning a reliable mapping between state features and failure probability. Without a clean separation between high-risk and low-risk states, calibration cannot provide a useful structure for adversarial decision-making and may even obscure the intended signal.

These observations indicate that the limitations we observe do not undermine the core strengths of our approach. The Hopper results, for example, reveal that although both methods eventually converge to similar maximal rewards, RC-RARL learns significantly faster during the early and mid stages of training, providing a clear advantage in sample efficiency and learning stability. Meanwhile, the high-noise results show that when failure dynamics are dominated by randomness rather than state-dependent structure, calibration naturally becomes less informative. Overall, risk calibration is most effective when failures reflect meaningful task dynamics and the adversary can exploit calibrated risk

to target truly informative perturbations. Understanding how to maintain its benefits in late-stage training or under environments with extremely noisy failure signals remains an important direction for future work.

6. Conclusion

In this work, we proposed Risk-Calibrated Robust Adversarial Reinforcement Learning (RC-RARL), a framework that augments adversarial training with a calibrated failure-risk model. By enabling the adversary to focus its perturbation on genuinely high-risk states, our approach reduces unnecessary disturbances and improves the overall quality of adversarial interactions.

RC-RARL demonstrates consistent improvements in early- and mid-training performance, yielding faster convergence, more stable learning dynamics, and higher rewards under moderate disturbance levels. Our results also show that probability calibration plays a crucial role in shaping adversarial behavior, allowing the adversary to adapt its strategy in a more targeted and risk-aware manner.

At the same time, our experiments reveal that performance gains diminish in very late training stages when failures become rare, and extremely high disturbance levels can obscure meaningful risk signals. These insights highlight both the promise and the current limitations of risk-calibrated adversarial training, and point toward richer risk models and more adaptive calibration strategies as promising directions for future work.

7. Future Work

Our findings suggest several promising directions for future exploration. First, a deeper analysis of how disturbance magnitude affects risk prediction quality would help clarify the conditions under which calibration remains informative. In particular, understanding how noise levels influence failure-label separability may reveal whether the risk predictor degrades smoothly or undergoes sharp transitions in extremely noisy regimes.

Second, the current risk estimator is a simple logistic regression model. While its simplicity facilitates fast online updates, its representational capacity is limited. Future work may consider more expressive risk predictors, including neural networks, sequence models that capture temporal dependencies, Bayesian models with uncertainty estimates, or domain-aware feature extractors tailored to specific control tasks. Such models may better capture complex failure boundaries and remain reliable even when noise levels increase.

Finally, exploring adaptive calibration strategies or dynamically adjusting the gating threshold could help maintain the benefits of risk-aware adversarial training throughout all stages of learning.

References

[1] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. "Robust adversarial reinforcement learning". In: *International conference on machine learning*. PMLR. 2017, pp. 2817–2826.

[2] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. "On calibration of modern neural networks". In: *International conference on machine learning*. PMLR. 2017, pp. 1321–1330.

[3] Ali Malik, Volodymyr Kuleshov, Jiaming Song, Danny Nemer, Harlan Seymour, and Stefano Ermon. "Calibrated model-based deep reinforcement learning". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 4314–4323.

[4] Yao Qin, Xuezhi Wang, Alex Beutel, and Ed Chi. "Improving calibration through the relationship with adversarial robustness". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 14358–14369.