ADVERSARIAL POLICIES: ATTACKING DEEP REINFORCEMENT LEARNING

IE708: Markov Decision Processes

Submitted by: Praharsh Agrawal (20D110014)

Submitted to: Prof. N. Hemachandra

Introduction

The research discusses the emergence of a new research field following the discovery of adversarial examples in image classifiers, expanding the investigation to deep reinforcement learning (RL) policies' vulnerability to adversarial perturbations.

Unlike image classifiers, RL agents operate in natural environments influenced by external agents, including humans, who can modify observations indirectly through actions.

The paper explores the feasibility of attacking RL policies by crafting adversarial policies that induce natural observations with adversarial effects on victims. The study uses simulated robotics games, demonstrating the existence of effective adversarial policies that outperform robust victims.

Interestingly, adversaries succeed not by becoming strong opponents but by manipulating observations through actions. The vulnerability is found to increase in higher-dimensional environments.

The paper proposes a physically realistic threat model for RL adversarial examples, emphasizes the importance of practitioner awareness, and suggests adversarial training as a tool for improving robustness in RL.

Objective

To investigate the vulnerability of RL policies to adversarial attacks and to analyze the impact of different hyperparameters and computational infrastructure on the effectiveness of attacks.

We demonstrate the existence of adversarial policies in zero-sum games between simulated humanoid robots with proprioceptive observations, against state-of-the-art victims trained via self-play to be robust to opponents. The adversarial policies reliably win against the victims but generate seemingly random and uncoordinated behavior. We find that these policies are more successful in high-dimensional environments, and induce substantially different activations in the victim policy network than when the victim plays against a normal opponent. Fine-tuning protects a victim against a specific adversary, but the attack method can be successfully reapplied to find a new adversarial policy.

Methodology

Adversarial policies were trained using Proximal Policy Optimization (PPO) algorithm to maximize the sum of discounted rewards. The training involves assigning sparse rewards at the end of episodes, with positive rewards when the adversary wins and negative rewards for losses or ties. This approach aligns with the reward structure used for training victim policies in prior work.

The adversarial policy is trained over 20 million timesteps using the PPO implementation from Stable Baselines. Hyperparameters are selected through a combination of manual tuning and a random search of 100 samples, as detailed in Section A of the appendix. The performance of the adversarial policy is compared to three baselines: a policy ‘**Rand**’ taking random actions, a lifeless policy ‘**Zero**’ with zero control, and pre-trained policies ‘**Zoo***’ from previous researches.

Victims were trained via self-play against random old versions of their opponent.

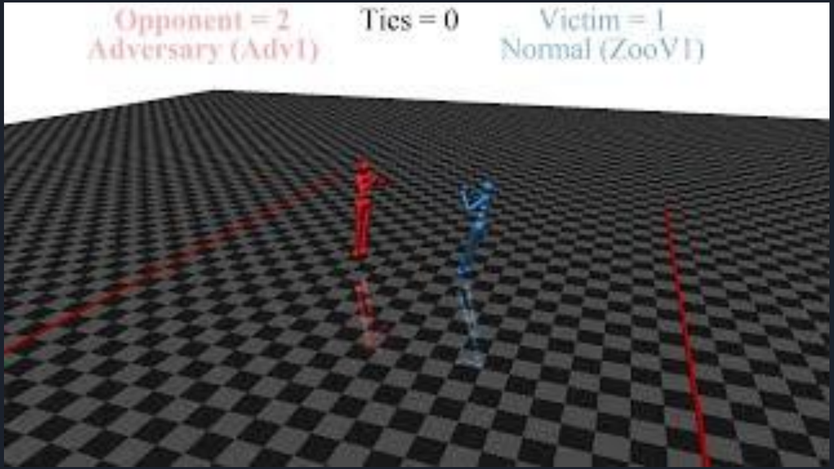
Four zero-sum simulated robotics games were used for evaluation: Kick and Defend, You Shall Not Pass, Sumo Humans, and Sumo Ants.

Results- Qualitative Evaluation

Adversarial policies exploit weaknesses in victim policies rather than performing the intended task.

Videos showcasing the strategies of adversarial policies are available for further understanding.

For example, in *Kick and Defend* and *You Shall Not Pass*, the adversarial policy never stands up. The adversary instead wins by positioning their body to induce adversarial observations that cause the victim's policy to take poor actions.



Distribution Shift

The effectiveness of adversarial policies against baselines suggests their robustness to off-distribution observations, while the resilience of victims to non-adversarially optimized inputs highlights the challenges and complexities in adversarial scenario exploration. This nuanced understanding is essential for advancing strategies that can secure machine learning systems against a broad range of potential threats.

To test this, we evaluate victims against two simple off-distribution baselines: a random policy ‘Rand’ (green) and a lifeless policy ‘Zero’ (red). These baselines win as often as 30% to 50% in *Kick and Defend*, but less than 1% of the time in *Sumo* and *You Shall Not Pass*. This is well below the performance of our adversarial policies.

Non-Transitive Relationships

This results are surprising as it implies highly non-transitive relationships may exist between policies even in games that seem to be transitive. A game is said to be transitive if policies can be ranked such that higher-ranked policies beat lower-ranked policies. Prima facie, the games in this paper seem transitive: professional human soccer players and sumo wrestlers can reliably beat amateurs. Despite this, there is a non-transitive relationship between adversarial policies, victims and masked victims. Consequently, we urge caution when using methods such as self-play that assume transitivity, and would recommend more general methods where practical.

The complexity arises from the dynamic and evolving nature of these interactions, challenging conventional expectations. This nuanced understanding is vital for effective adversarial defense strategies, urging comprehensive evaluations that go beyond direct adversarial impacts to consider potential indirect or concealed effects.

Trade-off in Observation Space

1. Performance and Observation Richness: - Enriching observations in RL boosts task performance, providing agents with more comprehensive data for improved decision-making.
2. Vulnerability to Adversarial Attacks: - Richer observations increase susceptibility to adversarial attacks, enabling adversaries to exploit detailed information for manipulation and potentially compromising system behavior.
3. Value of Information in RL Policies: - Not all additional information enhances RL policy performance. In certain contexts, extra details may not contribute positively, posing risks such as increased vulnerability to manipulation.
4. Trade-off and Decision-Making in RL: - RL necessitates a delicate balance between information richness and security. Striking the right equilibrium ensures optimal performance without compromising system integrity.
5. Implications for RL Policy Design: - Designing robust RL policies requires techniques like adversarial training and security-aware components to mitigate vulnerabilities associated with increased observation richness while maintaining performance gains.

Dimensionality

- Reinforcement learning policies are observed to be more susceptible to adversarial attacks when operating in high-dimensional input spaces, as per the findings. In such scenarios, where input data is characterized by a multitude of features, the policies are more prone to misbehavior induced by intentionally crafted perturbations.
- The results substantiate the hypothesis that increased dimensionality heightens the vulnerability of reinforcement learning models to adversarial manipulation. This implies that the intricate and complex nature of high-dimensional input spaces poses challenges for the model to discern genuine patterns from adversarial influences.
- Recognizing and addressing this susceptibility is crucial for the development of robust reinforcement learning systems.
- Strategies like adversarial training and robust optimization become imperative to mitigate these vulnerabilities, ensuring the responsible and secure deployment of reinforcement learning in diverse and complex real-world environments.

Hyperparameter Selection (t-SNE)

The impact of varying perplexity values, specifically 5, 10, 20, 50, 75, 100, 250, and 1000, on data visualization in the context of probabilistic models. Notably, a perplexity of 250 emerged as optimal, yielding the clearest representation of the data with a moderate number of distinct clusters. Perplexity is a critical hyperparameter in models like t-SNE, influencing the balance between preserving intricate structures and avoiding over-segmentation. The choice of 250 struck a harmonious balance, leading to a visualization that effectively captured underlying patterns while maintaining interpretability. This finding suggests the importance of thoughtful perplexity selection in dimensionality reduction techniques, offering insights into how parameter tuning can enhance the clarity and meaningful representation of complex datasets.

Figures

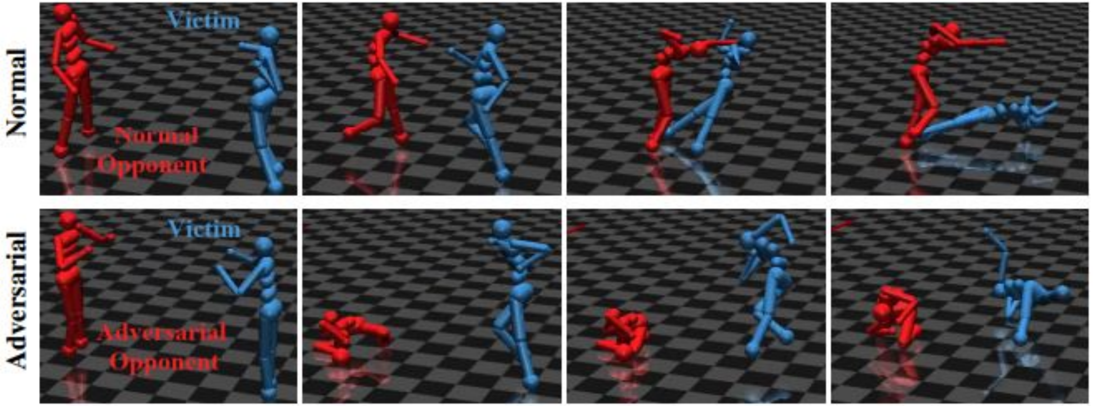


Figure 1: Illustrative snapshots of a victim (in blue) against normal and adversarial opponents (in red). The victim wins if it crosses the finish line; otherwise, the opponent wins. Despite never standing up, the adversarial opponent wins 86% of episodes, far above the normal opponent's 47% win rate.

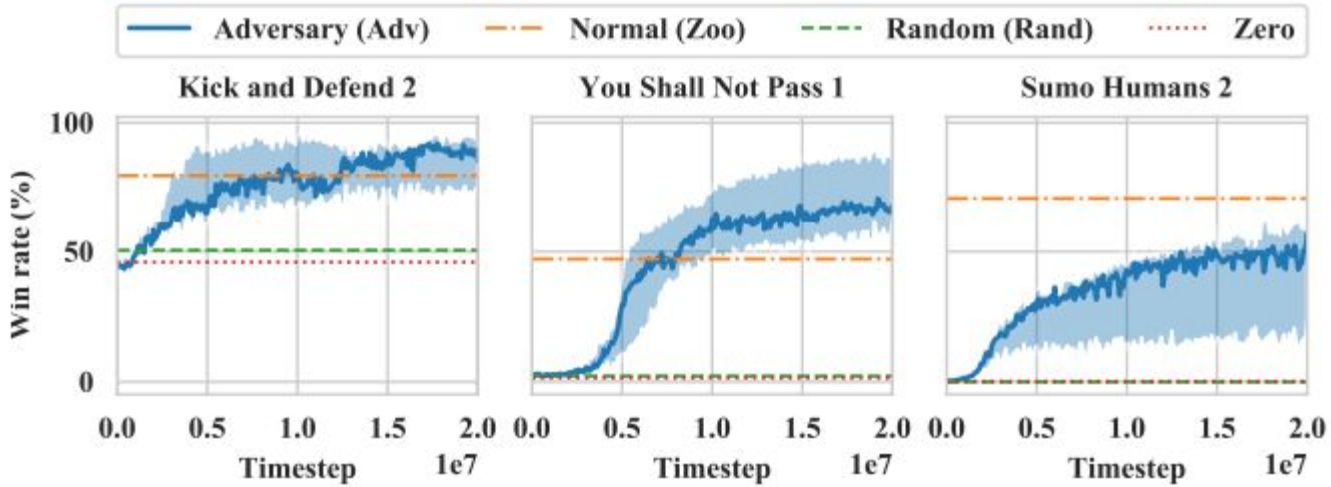
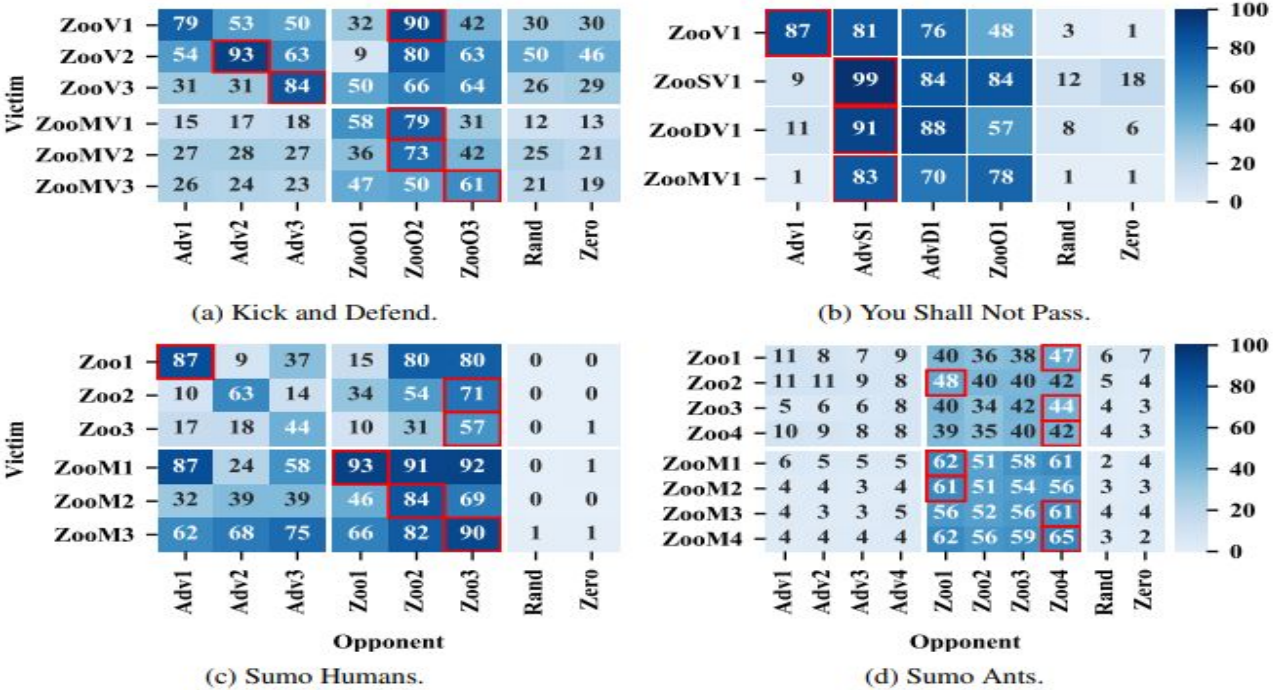


Figure 3: Win rates while training adversary Adv against the median victim in each environment (based on the difference between the win rate for Adv and Zoo). The adversary outperforms the Zoo baseline against the median victim in *Kick and Defend* and *You Shall Not Pass*, and is competitive on *Sumo Humans*. For full results, see figure 4 below or figure C.1 in the supplementary material.



Victim	Adv1	Adv2	Adv3	ZooO1	ZooO2	ZooO3	Rand	Zero
ZooV1	79	53	50	32	90	42	30	30
ZooV2	54	93	63	9	80	63	50	46
ZooV3	31	31	84	50	66	64	26	29
ZooMV1	15	17	18	58	79	31	12	13
ZooMV2	27	28	27	36	73	42	25	21
ZooMV3	26	24	23	47	50	61	21	19

(a) Kick and Defend.

	Adv1	AdvS1	AdvD1	ZooO1	Rand	Zero
ZooV1	87	81	76	48	3	1
ZooSV1	9	99	84	84	12	18
ZooDV1	11	91	88	57	8	6
ZooMV1	1	83	70	78	1	1

(b) You Shall Not Pass.

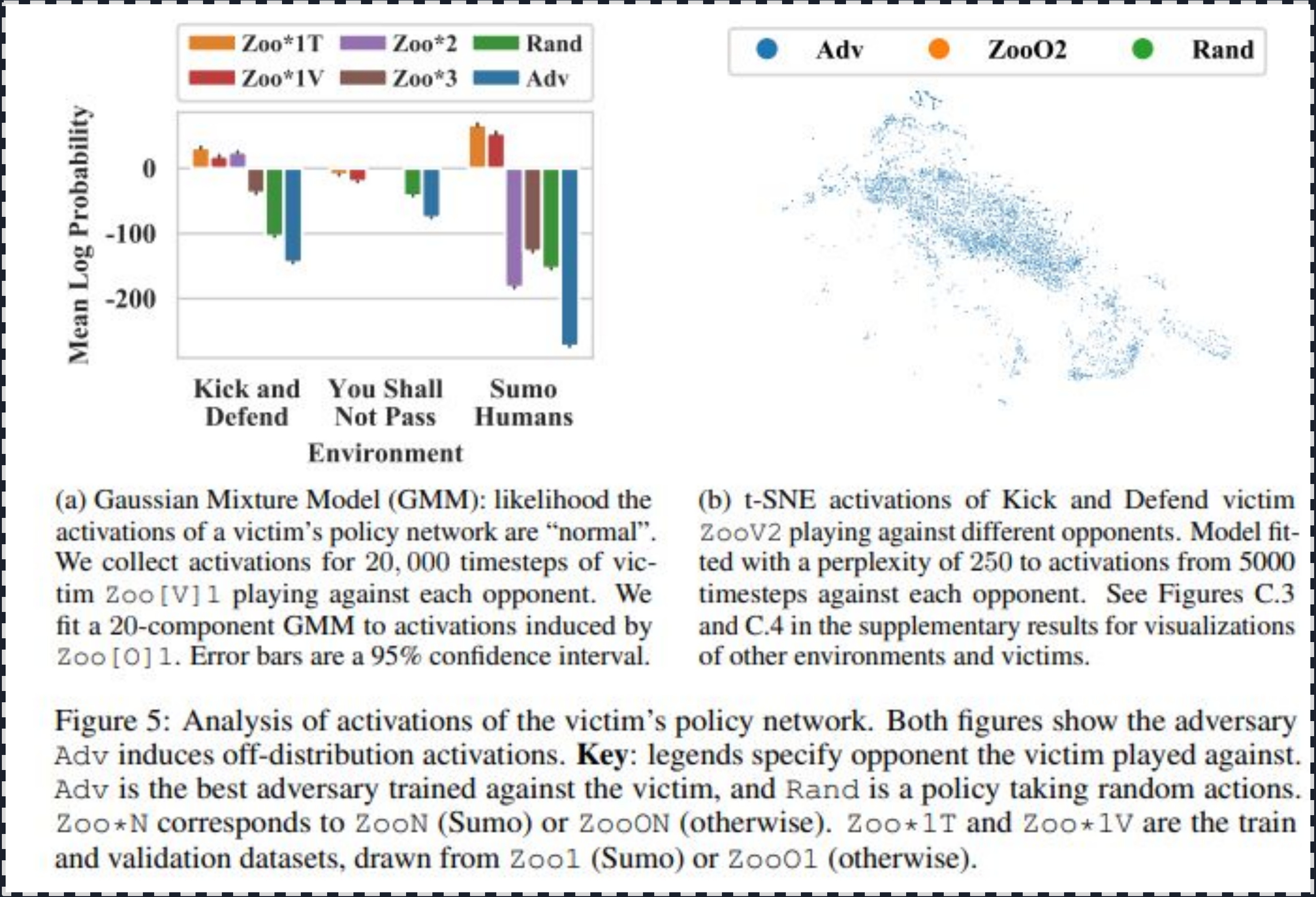
Victim	Adv1	Adv2	Adv3	Zoo1	Zoo2	Zoo3	Rand	Zero
Zoo1	87	9	37	15	80	80	0	0
Zoo2	10	63	14	34	54	71	0	0
Zoo3	17	18	44	10	31	57	0	1
ZooM1	87	24	58	93	91	92	0	1
ZooM2	32	39	39	46	84	69	0	0
ZooM3	62	68	75	66	82	90	1	1

(c) Sumo Humans.

	Adv1	Adv2	Adv3	Adv4	Zoo1	Zoo2	Zoo3	Zoo4	Rand	Zero
Zoo1	11	8	7	9	40	36	38	47	6	7
Zoo2	11	11	9	8	48	40	40	42	5	4
Zoo3	5	6	6	8	40	34	42	44	4	3
Zoo4	10	9	8	8	39	35	40	42	4	3
ZooM1	6	5	5	5	62	51	58	61	2	4
ZooM2	4	4	3	4	61	51	54	56	3	3
ZooM3	4	3	3	5	56	52	56	61	4	4
ZooM4	4	4	4	4	62	56	59	65	3	2

(d) Sumo Ants.

Figure 4: Percentage of games won by opponent (out of 1000), the maximal cell in each row is in red. **Key**: Agents ZooYN are pre-trained policies from Bansal et al. (2018a), where $Y \in \{‘V’, ‘O’, ‘’\}$ denotes the agent plays as **(V)**ictim, **(O)**pponent or either side, and N is a random seed. Opponents AdvN are the best adversarial policy of 5 seeds trained against the corresponding Zoo[V]N. Agents Rand and Zero are baseline agents taking random and zero actions respectively. Defended victims ZooXYN, where $X \in \{‘S’, ‘D’, ‘M’\}$, are derived from ZooYN by fine-tuning against a **(S)**ingle opponent AdvN, **(D)**ual opponents AdvN and Zoo[O]N, or by **(M)**asking the observations.



(a) Gaussian Mixture Model (GMM): likelihood the activations of a victim's policy network are "normal". We collect activations for 20,000 timesteps of victim `Zoo[V]1` playing against each opponent. We fit a 20-component GMM to activations induced by `Zoo[O]1`. Error bars are a 95% confidence interval.

(b) t-SNE activations of Kick and Defend victim `ZooV2` playing against different opponents. Model fitted with a perplexity of 250 to activations from 5000 timesteps against each opponent. See Figures C.3 and C.4 in the supplementary results for visualizations of other environments and victims.

Figure 5: Analysis of activations of the victim's policy network. Both figures show the adversary `Adv` induces off-distribution activations. **Key**: legends specify opponent the victim played against. `Adv` is the best adversary trained against the victim, and `Rand` is a policy taking random actions. `Zoo*N` corresponds to `ZooN` (Sumo) or `ZooON` (otherwise). `Zoo*1T` and `Zoo*1V` are the train and validation datasets, drawn from `Zoo1` (Sumo) or `ZooO1` (otherwise).

Observation Space and Vulnerability

RL policies show a trade-off between observation space size and vulnerability to adversarial attacks. More observation of the environment improves performance but increases susceptibility (or vulnerability) to adversaries with intentional manipulations.

Adversarial attacks can exploit the complexity of extensive observations, posing a dilemma where optimizing for superior performance may compromise the model's security. Striking the right balance is crucial, requiring RL practitioners to carefully assess the trade-off based on the specific needs of the application.

The findings underscore the need for nuanced approaches in RL design, particularly in real-world scenarios like autonomous driving or financial trading, where adversaries may exist. Careful consideration of observation space design is essential to ensure both optimal performance and resilience against potential adversarial threats.

Non-Transitivity in Adversarial Policies

Adversarial policies win not by physically interfering with victims but by strategically inducing natural observations that are adversarial. To test this, a 'masked' victim is introduced, where the observation of the adversary's position is fixed to a static value. Surprisingly, despite the expectation that seeing the opponent is beneficial, the masked victims perform worse against normal opponents.

However, when playing against adversaries, the normal victims fare poorly compared to the masked victims. This reveals non-transitive relationships among adversarial policies, victims, and masked victims, challenging assumptions of transitivity even in seemingly transitive games.

The findings caution against relying on self-play methods that assume transitivity and suggest a trade-off between observation space size, performance, and vulnerability to adversaries in RL environments. The next section delves into investigating this connection further.

Dimensionality and Vulnerability

A variety of work has concluded, derived from prior research, that classifiers are more vulnerable to adversarial examples on high-dimensional inputs.

The research hypothesizes a similar result for RL policies: the greater the dimensionality of the component P of the observation space under control of the adversary, the more vulnerable the victim is to attack.

We test this hypothesis in the Sumo environment, varying whether the agents are Ants or Humanoids. The results in Figures 4c and 4d support the hypothesis. The adversary has a much lower win-rate in the low-dimensional Sumo Ants (dim P = 15) environment than in the higher dimensional Sumo Humans (dim P = 24) environment, where P is the position of the adversary's joints.

Conclusions

The research underscores the pervasive success of adversarial policies in undermining deep reinforcement learning (RL) policies within zero-sum simulated robotics games.

This highlights the intrinsic vulnerabilities present in RL systems and emphasizes the critical need for a profound understanding of these weaknesses.

The delicate balance between observation space size and susceptibility to adversarial attacks emerges as a central theme, demanding careful consideration in RL policy design for achieving both high performance and enhanced security.

Moreover, the identification of non-transitive relationships challenges conventional assumptions in RL, urging caution in methodologies that rely on transitivity and advocating for more versatile approaches.

Conclusions

The observed sensitivity of RL policies to the dimensionality of the observation space further accentuates the complexity of designing resilient RL systems.

As RL finds increasing applications in real-world environments, the study emphasizes the importance of practitioners being cognizant of the threat model posed by adversarial policies.

Notably, the potential of adversarial training using adversarial policies is highlighted as a promising avenue for fortifying RL systems against a spectrum of potential weaknesses, showcasing the dynamic and evolving nature of adversarial dynamics in RL research.

Future Research

Overall the research is exciting because of the implications that the adversarial policy model has, for the robustness, security and understanding of deep RL policies.

Investigate the effectiveness of different defense mechanisms against adversarial attacks. Explore the impact of adversarial attacks on RL policies in real-world scenarios. Some of the problems that could be solved in the future using this research are:

- Robustness Enhancement in AI Systems
- Improved Security Measures
- Adaptive Learning Strategies
- Enhanced Autonomous Systems
- Ethical Considerations and Responsible AI

References

1. https://axrp.net/episode/2020/12/11/episode-1-adversarial-policies-adam-gleave.html
2. https://www.youtube.com/watch?v=8OkkBaPazl8&ab_channel=AXRP
3. https://www.gleave.me/
4. ADVERSARIAL POLICIES: ATTACKING DEEP REINFORCEMENT LEARNING by Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, Stuart Russell (UC Berkley)
5. ChatGPT (GPT 3.5) for summarising text and gaining some additional insights

THANK YOU