The University of Texas Rio Grande Valley
Center for Multidisciplinary Research Excellence in Cyber-Physical Infrastructure Systems (MECIS)

Cybersecurity of AI-powered Traffic Signal Control

National Science Foundation
NSF Award No. 2112650

Mark Hernandez, Mohamadhossein Noruzoliaee, Ph.D., Fatemeh Nazari, Ph.D.

Abstract

Next-generation transportation systems are increasingly integrated cyber (e.g., AI) and physical (e.g., traffic signals) systems, which exposes these systems to high-risk cyber threats, potentially causing service disruptions and economic losses. Many AI-powered cyber-physical transportation systems leverage reinforcement learning (RL) for traffic control and optimization, but RL has been recently found to be intrinsically vulnerable to cyberattacks. To tackle, a game-theoretic adversarial cyber-defense model is proposed that utilizes RL to learn an optimal adversarial policy to build a certifiably robust agent in the traffic control setting under complex, hybrid attacks. The proposed approach aims to certify the security of AI-powered transportation systems under evolving cybersecurity threats.

Introduction & Background

- Emerging cybersecurity threats in RL-driven traffic systems arise from adversarial attacks on agent-environment dynamics
- Adversarial vulnerabilities in RL-driven traffic signal control pose risks such as high economic losses due to transportation network-wide congestion

PROBLEM

- Adversarial agents exploit RL policies by introducing perturbations into the agent-environment interaction, compromising performance, and exposing cybersecurity risks

GOAL

- Enhance RL policy robustness against cybersecurity threats by integrating learned adversarial models into the training process

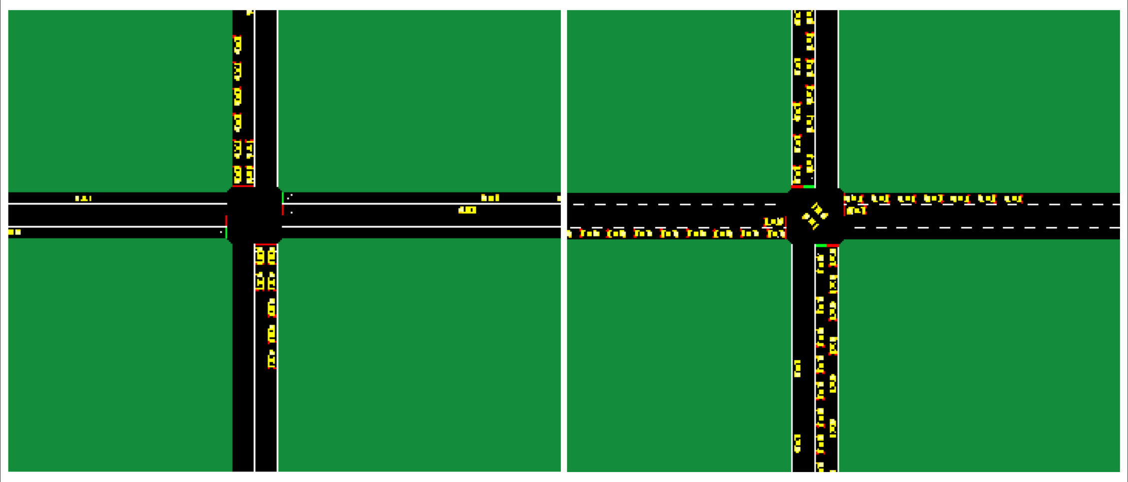

Figure 1: (Left) Traffic Signal Control via RL Agent, (Right) RL Agent Under Attack

Methodology

- Hybrid adversarial model perturbs both state and action
  - *State attack:* changes in perceived traffic data from environment
  - *Action attack:* manipulates agent-selected signal phase
- Alternating agent and adversary training using ATLA
  - Adversary learns perturbations that maximize agent reward penalties
  - Agent learns to adapt for robustness
- Agent and adversary implemented with neural networks, optimized via Proximal Policy Optimization

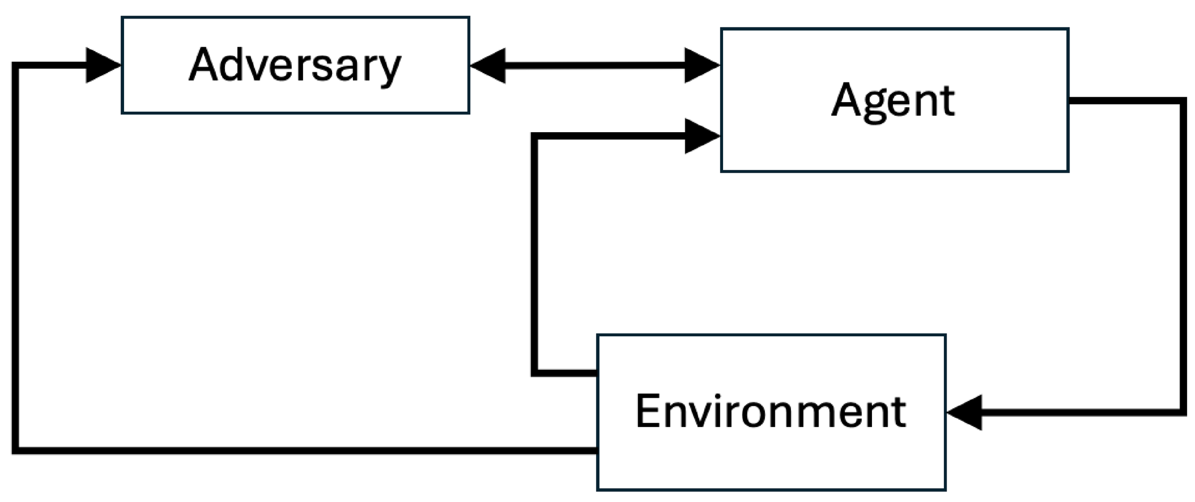

Figure 2: Agent-Adversary Zero–Sum Game Framework

Data and Results

- *Reward:* calculated using accumulated waiting time per lane
- *State:* phase identifier, lane densities, lane queues, minimum green time indicator
- *Action:* four possible green phase configurations each followed by yellow phase

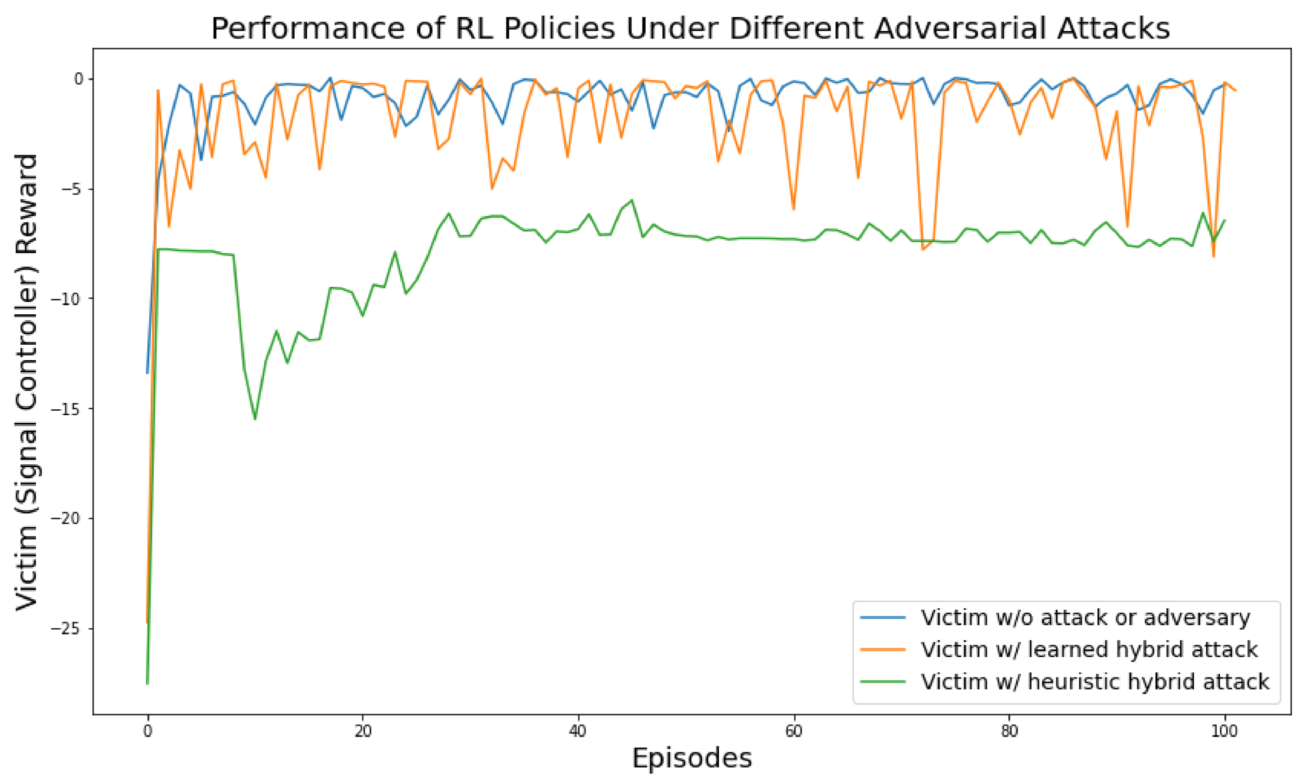

Figure 3: Robustness Gain vs Performance Loss

Table 1: Normalized Reward for Different Attacks and Traffic Volumes

Attack Type on Victim (Signal Controller)	Average Reward by Traffic Volume		
	Low	Moderate	High
No Attack	-0.865	-1.185	-3.016
Learned Hybrid Attack	-1.652	-1.869	-3.105
Heuristic Hybrid Attack	-7.199	-8.002	-7.248

Conclusions & Future Work

- Evaluated baseline heuristic hybrid evasion attacks on a victim agent that simultaneously perturbs state observations and actions
- Proposed the integration of the ATLA framework into RL-driven traffic signal control to enhance policy robustness
- Demonstrated the effectiveness of the ATLA framework in mitigating cybersecurity threats and providing certifiable robustness through a learned hybrid attack

Acknowledgments

We acknowledge the financial support provided by the NSF CREST Center for Multidisciplinary Research Excellence in Cyber-Physical Infrastructure Systems (MECIS) through NSF Award No. 2112650.

References

[1] Sun, Yanchao, Ruijie Zheng, Yongyuan Liang, and Furong Huang. "Who is the strongest enemy? towards optimal and efficient evasion attacks in deep rl." *arXiv preprint arXiv:2106.05087* (2021).

[2] Zhang, Huan, Hongge Chen, Duane Boning, and Cho-Jui Hsieh. "Robust reinforcement learning on state observations with learned optimal adversary." *arXiv preprint arXiv:2101.08452* (2021).

[3] LucasAlegre. sumo-rl. Github. Last modified September 4, 2024. https://github.com/LucasAlegre/sumo-rl