Securing Dynamic Robotic Behavior in Unpredicted Environments: Enhancing Trust through Adaptive Learning and Cyber Defense

Giselle Roman | Dr. Yugyung Lee | 06/16/2025

Problem Statement Overview

Problem: Traditional robots are unable to handle unpredictable, dynamic environments. Learning-based methods (RL/IL) help improve adaptability but raises concern with trust and security.

Goal: Create an adaptive and secure robotic systems using RL with cyber defense mechanisms.

Hypothesis & Research Questions

Hypotheses: Incorporating adversarial input detection into reinforcement learning training will result in lower collision rates during mapless navigation.

Research Questions:

1. What is the most effective way to reduce the collision rate in mapless navigation without interfering with exploration and adaptability?
2. How can mitigating adversarial attacks during reinforcement learning reduce the risk of external manipulation during operation?

Paper 1: A Deep Safe Reinforcement Learning Approach

- Safe robot navigation in unknown environments using DRL
  - Mapless navigation as a scalable alternative
- Safe DRL framework
  - Constrained Policy Optimization (CPO): to optimize navigation under safety constraints
  - Actor-Critic Safety (ACS) Architecture: to manage reward and risk during training
- Reframed collisions as negative rewards
  - Risk-taker
  - Risk-seeker

Paper 2: Collaborative Assembly in Hybrid Manufacturing Cells

- Trust-aware Human-Robot Collaboration (HRC)
  - Physical and social interaction in manufacturing environments
- Physical HRI (pHRI) + social HRI (sHRI)
  - Improve coordination and reduce human workload
- Trust model
  - Facial emotion recognition
    - Adjust to human pace
    - Display emotional feedback
- Real-world trials
  - 44% reduction in human workload

Paper 3: Securing Cyber-Physical Robotic Systems

- Cyber-physical robotic systems (CPRS)
  - Physical sensors and digital channels
  - Detect adversarial behavior
- Robotic safety
  - Real-time monitoring
  - Attack classification
- Layered CIAAP taxonomy
  - Map robotic vulnerabilities
- Attack tree simulation model
  - Analyze known and unknown threats
  - Preemptive defenses

AI Methods

- Constrained Policy Optimization (CPO)
  - For safety during navigation
- Autoencoder
  - For anomaly detection in sensor feedback
- Graph Neural Networks (GNNs)
  - To model spatial relations and trust mapping

Challenges

- Transitions from simulation to real world environments
- Trust feedback system
  - Measure and responding to users
- Implementing Runtime Adversarial Defenses
  - Best way to detect or defend against adversarial inputs
- System Integration
  - Learning models, cybersecurity tools, and trust

Next Steps

Week	Focus
Week 3 (June 10–14)	- Refine RL/IL training - Begin adversarial testing - Trust scoring exploration
Week 4 (June 17–21)	- Build trust feedback loop - Test adversarial behavior - Upload progress to GitHub
Week 5 (June 24–28)	- Add basic transparency - Continue code activity
Week 6 (July 1–5)	- Run full simulations - Continue to work on report and slides
Week 7 (July 8–12)	- Finalize experiments - Test and validate results
Week 8 (July 15–19)	- Revise report/slides - Organize GitHub
Week 9 (July 22–26)	- Practice presentation - Final cleanup
Final Week (July 29–31)	- Give final presentation

References

- Bhardwaj, Akashdeep, Salil Bharany, Ateeq Ur Rehman, Ghanshyam G. Tejani, and Seada Hussen. “Securing Cyber-Physical Robotic Systems for Enhanced Data Security and Real-Time Threat Mitigation.” EURASIP Journal on Information Security, vol. 2025, no. 1, Jan. 2025, https://doi.org/10.1186/s13635-025-00186-7.
- Lv, Shaohua, Yanjie Li, Qi Liu, Jianqi Gao, Xizheng Pang, and Meiling Chen. “A Deep Safe Reinforcement Learning Approach for Mapless Navigation.” 2021 IEEE International Conference on Robotics and Biomimetics (ROBIO), Dec. 2021, pp. 1520–25. IEEE, https://doi.org/10.1109/ROBIO54168.2021.9739251.
- Sadrfaridpour, Behzad, and Yue Wang. “Collaborative Assembly in Hybrid Manufacturing Cells: An Integrated Framework for Human–Robot Interaction.” IEEE Transactions on Automation Science and Engineering, vol. 15, no. 3, July 2017, pp. 1178–92, https://doi.org/10.1109/tase.2017.2748386.