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

Giselle Roman | Dr. Yugyung Lee | 06/08/2025

Problem Statement Overview

As autonomous robots continue to operate in complex and unpredictable environments, their ability to adapt in real-time scenarios is important. By exploring learning-based methods like reinforcement and imitation learning, this project aims to optimize robotic behavior by developing both adaptive learning and highly secure cybersecurity.

Paper 1: Secure Robotics: Safety, Trust, and Cybersecurity

- Introduced the unified field of secure robotics
  - Trust
  - Safety
  - Cybersecurity
- Trust as non-binary, shaped over time based on experience
  - Asimov's laws
    - Human values and ethics
- Provided taxonomy of trust failures
  - System Failure
  - User Failure

Paper 2: AI and ML Enhance Robot Decision-making

- Survey-style overview of AI/ML in robotics
- Emphasized real-time adaptation
  - Reinforcement Learning (RL): exploring its environment to receive feedback
  - Intimidation Learning (IL): data provided to learn from
- Showed case studies
  - Da Vinci Surgical System
  - Boston Dynamics' Spot Robot

Paper 3: Adversarial Attacks

- Deep Reinforcement Learning (DRL) agents
  - Vulnerable to small perturbations
- Focused on white-box vs black-box attacks
  - White-box: attacker has full access to the robot's model
  - Black-box: attacker has no internal knowledge
- Fast Gradient Sign Method (FGSM) attack method
  - Adversarial attacks during training and testing

AI Methods

- Reinforcement Learning (RL)
  - Adaptive decision-making
- Imitation Learning (IL)
  - Learning from demonstrations
- Adversarial Robustness
  - Attack simulations to test defenses
- Perspective API
  - Potential tool for modeling trust and feedback

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

- Haskard, Adam, and Damith Herath. “Secure Robotics: Navigating Challenges at the Nexus of Safety, Trust, and Cybersecurity in Cyber-Physical Systems.” *ACM Computing Surveys*, Association for Computing Machinery, Mar. 2025, https://doi.org/10.1145/3723050.
- Huang, Sandy, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial Attacks on Neural Network Policies. ArXiv:1702.02284 [cs, stat], Feb. 2017, https://arxiv.org/abs/1702.02284.
- Hussain, Md Delwar, Md Hamidur Rahman, and Nur Mohammad Ali. “Artificial Intelligence and Machine Learning Enhance Robot Decision-Making Adaptability and Learning Capabilities across Various Domains.” ResearchGate, unknown publisher, 3 June 2024, www.researchgate.net/publication/381131067_ARTIFICIAL_INTELLIGENCE_AND_MACHINE_LEARNING_ENHANCE_ROBOT_DECISION-MAKING_ADAPTABILITY_AND_LEARNING_CAPABILITIES_ACROSS_VARIOUS_DOMAINS.