Securing Dynamic Robotic Coordination in Unpredictable Environments: Enhancing Trust through Adaptive Learning and Cyber Defense

Giselle Roman
Department of Computer Science
California State University, Fullerton

This work was conducted as part of the NSF Research Experiences for Undergraduates (REU) Program at the University of Missouri–Kansas City, Summer 2025.

Abstract: Autonomous robotic systems deployed in unpredictable and disaster-prone environments face significant cybersecurity challenges that can degrade trust and disrupt coordinated operations. This work models adversarial attacks that systematically poison trust relationships between agents, reducing the effectiveness of collaborative task allocation. To address this, a simulation framework is developed that incorporates randomized fleet trust degradation and proportional defense countermeasures that restore degraded trust scores. Task assignment optimization under these trust constraints is formulated as a Quadratic Unconstrained Binary Optimization (QUBO) problem, solved using a NEAL sampler. Results demonstrate that while adversarial attacks significantly lower trust-based task assignment quality, adaptive defense strategies partially restore system coordination, offering a viable path for resilient multiagent planning in adversarial settings.

Index Terms: Cybersecurity, Trust Modeling, Multi-Agent Systems, Reinforcement Learning, Quadratic Unconstrained Binary Optimization (QUBO), Robotic Coordination

1. Introduction

In post-disaster scenarios where environments are too dangerous or unstable for human responders, autonomous robotic systems can be deployed to navigate rubble, inspect damaged structures, and complete critical rescue missions. These robots, like quadruped robots or drones, operate in unpredictable environments while maintaining reliable coordination to achieve mission goals.

However, as these systems become more capable and adaptive, they also become increasingly vulnerable to cyber threats. Adversarial attacks such as GPS spoofing and sensor noise injection can degrade inter-agent trust and disrupt collaborative task allocation. Although there are defensive strategies that incorporate penalties, rewards, and encouragement mechanisms, they must be carefully designed to prioritize the tasks that are most urgent, which agents are best suited for them, and how trust in these agents can be managed, especially when trust has been compromised.

This research addresses these challenges by modeling a fleet of autonomous agents that can experience systematic trust degradation from attacks and partial trust restoration through defense strategies. Each agent's trust score represents its reliability for different tasks, which changes dynamically under attack and defense conditions. To optimize task assignments under various trust constraints, the problem is designed using a Quadratic Unconstrained Binary Optimization (QUBO) model. The QUBO formulation allows for deciding which agent should be assigned to which task by maximizing overall trust scores, even when those scores have been partially poisoned. This

is decided through QUBO using a NEAL sampler to explore high-quality assignment solutions in realistic adversarial scenarios.

2. Related Work

Haskard and Herath's work on *Secure Robotics: Navigating Challenges at the Nexus of Safety, Trust, and Cybersecurity in Cyber-Physical Systems* highlights the critical need to address safety, trust, and cybersecurity together in robotic systems. Their discussion motivates modeling adversarial attacks that can degrade trust between agents, a central challenge this work addresses by simulating trust degradation and evaluating defense strategies.

Bojchevski and Günnemann's study on *Adversarial Attacks on Node Embeddings via Graph Poisoning* demonstrates how adversarial perturbations can systematically degrade graph-based representations. Inspired by this approach, this project treats the agent-task trust matrix as a graph-like structure and simulates poisoning attacks to evaluate how degraded trust impacts task assignment quality, and how defense strategies can partially recover system performance.

3. Methodology

Models a post-disaster rescue scenario where teams of autonomous robots must assign tasks while under potential adversarial attack. The system pipeline is depicted in the conceptual diagram: Initialization $\rightarrow$ Attack $\rightarrow$ Defense $\rightarrow$ Optimization $\rightarrow$ Evaluation.

4. Experimental Results and Evaluation

Conducted simulation experiments to evaluate the effect of adversarial attacks and defenses on trust-aware task assignment.

4.1. Input Trust Matrices

Generated three trust matrix scenarios:

- **Baseline:** Initial trust without attacks. Higher values indicate agent confidence.
- **Poisoned:** Trust degraded system-wide to simulate attack. Lower values reduce assignment quality.
- **Defended:** Countermeasure logic partially restores trust scores below a threshold.

4.2. Assignment Matrices (QUBO Outputs)

QUBO optimization generates binary assignment matrices, where each 1 indicates an agent-task assignment under trust constraints.

- **Baseline Assignment:** typically better spread, few violations
- **Poisoned Assignment:** degraded trust leads to suboptimal assignments
- **Defended Assignment:** partial recovery improves feasibility but may still show some constraint violations

4.3. Heatmaps

Visualize trust matrices and overlay QUBO assignments:

- Color scale represents trust score magnitudes.
- Grey circles mark QUBO-selected assignments.

Preliminary Findings:

- Attacks significantly lower trust and degrade assignment quality.
- Simple defenses improve trust but don't always restore baseline performance.
- Careful tuning of countermeasures and penalty terms is needed to enforce one-to-one constraints.

5. Conclusions

Explores the challenge of securing multiagent robotic coordination in adversarial environments where trust between agents can be systematically degraded by cyber attacks. By simulating randomized trust degradation and applying partial defense countermeasures, we model realistic scenarios where trust scores impact the feasibility of task assignment.

6. References

...

Acknowledgements

This work was supported by the National Science Foundation under the REU program. The author thanks Dr. Yugyung Lee for mentoring and guidance throughout the project.