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
gisellejbroman@csu.fullerton.edu

***Abstract*—With the advancement of robotic systems being deployed in unpredictable and unsafe environments, they face significant cybersecurity challenges. This degrades trust and disrupts coordinated operations with other robotic systems and humans. These simulations model adversarial attacks that poison trust between agents, reducing collaborative task allocation. By focusing on these issues, the simulations incorporate randomized trust degradation and proportional defense countermeasures to help partially restore trust scores. To optimize task assignments under trust constraints, Quadratic Unconstrained Binary Optimization (QUBO) was formulated and solved using a NEAL sampler. The results demonstrate that while adversarial attacks lower trust in task assignment, the defense countermeasures can partially restore trust, allowing for better multi-agent planning in adversarial settings.**

***Index Terms*—Cybersecurity, Trust Modeling, Multi-Agent Systems, Adversarial Attacks, Anomaly Detection, Defense Countermeasures, Reinforcement Learning, Quadratic Unconstrained Binary Optimization (QUBO), Robotic Coordination**

I. Introduction

In a post-disaster environment, conditions become too dangerous for human responders. This enables the deployment of robotic systems to navigate unstable terrain and complete critical rescue missions. Quadruped robots or drones can operate in these types of environments while maintaining collaboration to achieve mission success.

However, as these systems become more adaptable, they also become more vulnerable to cyber threats. Attacks such as GPS spoofing and sensor noise poisoning can degrade inter-agent trust and disrupt collaboration. Current defensive strategies integrate penalties, rewards, and encouragement-based countermeasures. These focus on prioritizing urgent tasks, identifying which agents are suitable for them, and managing trust scores in agents even when they are compromised.

This research models a fleet of agents that experience systematic trust degradation from attacks along with partial trust restoration through defense. Each agent has a trust score representing its reliability for different tasks. This score changes dynamically under attack and when defense countermeasures are applied. Under these constraints, task assignments are optimized using a QUBO model. QUBO decides which agent should be assigned by maximizing overall trust scores. By using a NEAL sampler, the model explores high-quality assignment solutions in realistic adversarial scenarios.

II. Related Work

Ensuring secure and resilient coordination between robotic systems in adversarial environments is a challenging task. Prior studies discuss the feasibility of attacks on graph-based representations for trust and the importance of defense strategies.

Graph-based learning methods like GCN and GAT are vulnerable to poisoning attacks [1]. By carefully adding or removing a small number of edges, attackers can degrade node embeddings without detection. Meta-learning-based attacks that target overall embedding quality or specific nodes show high transferability across models. This work treats the agent task trust matrix as graph-like structures that can be poisoned. This motivates the need for defense countermeasures that can detect and mitigate degradation in real-time task allocation.

The study of adversarial attacks on node embeddings [1] demonstrates how small, well-designed perturbations to a graph can significantly degrade embedding quality and downstream task performance. Their work shows that attacks are effective even with limited attacker knowledge and budget, and that defenses need to address structural vulnerabilities in graph representations. These insights highlight why robust trust modeling in multi-agent planning is essential, since our approach uses graph-like trust matrices that can be similarly attacked.

By integrating security and trust modeling, researchers have explored the intersection of cybersecurity, safety, and trust in cyber-physical systems [2]. This emphasizes that trust is not a binary concept but emerges through reliable and transparent interaction. A taxonomy of trust-relevant failures—including design, system, expectation, and user failures—explains how trust affects technical attacks and poor system designs. This motivates our research focus on real-time trust monitoring and the reallocation of tasks when anomalies are detected.

The concept of "secure robotics" [2] positions safety, trust, and cybersecurity as tightly connected elements that must be considered together when deploying robots in human environments. Their framework argues for integrating cybersecurity into system design from the start, not as an afterthought. This

is relevant to our approach, which simulates trust degradation under attack while incorporating defense strategies that aim to restore trust dynamically. Their emphasis on proactive security and maintaining user trust aligns with our goal of developing resilient task allocation under adversarial conditions.

These studies motivate the central hypothesis that integrating anomaly detection, trust modeling, and quantum-inspired optimization can create better multi-agent collaboration under adversarial conditions.

III. Methodology

This work models a fleet of dynamic robots that coordinate tasks in an adversarial environment where cyber-physical attacks can lower trust scores. The simulation pipeline includes trust initialization, simulated attacks, defense countermeasures, and task assignment optimization using a QUBO formulation.

A. Architecture Overview

Each agent is modeled with a trust score that indicates its reliability for completing assigned tasks. Agents are initialized with random trust values sampled from a realistic distribution to simulate fleet heterogeneity. These simulations vary the number of agents (50–100) and the number of rounds (100–1000) to explore different behaviors.

Across multiple simulation versions (v2, v3, v3.5), the architecture remains consistent, with agents' trust scores evolving dynamically over time due to attacks, defense countermeasures, and task performance feedback.

B. Attack Simulation

These adversarial attacks model systematic degradation of trust scores. Each attack introduces randomized noise subtracted from the trust matrix, lowering trust scores to represent realistic cyber threats such as GPS spoofing or communication interference.

The **Poisoned Trust Matrix** describes degraded reliability across the fleet, which impacts the quality of task assignments.

C. Defense Countermeasures

To defend against these attacks, the system applies proportional defense strategies that partially restore degraded trust scores. These parameters increase trust values for agents showing consistent task success while applying penalties for repeated failures.

This adaptive behavior models real-world cyber defense measures for anomaly detection and trust encouragement based on observed behavior.

D. QUBO Formulation for Task Assignment

For task allocation, trust variability is modeled using a QUBO formulation with the goal of maximizing total trust reward while satisfying hard assignment constraints:

- Each agent is assigned to exactly one task.
- Each task is assigned to exactly one agent.

The QUBO objective function includes:

- A trust reward term that encourages high-trust agent-task pairings.
- Hard constraint penalty terms enforcing assignment feasibility.
- An encouragement parameter that biases assignments toward more reliable agents.

The QUBO is solved using a NEAL simulated annealing sampler, which produces high-quality solutions even under degraded trust conditions.

E. System Pipeline

Figure 1 illustrates the overall simulation loop and system pipeline:

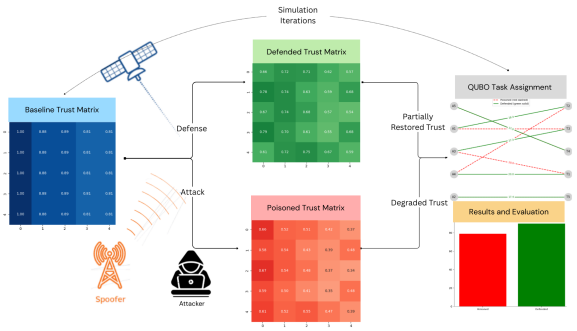

Fig. 1. Simulation pipeline illustrating attack degradation, defense countermeasures, and QUBO-based task assignment optimization.

- **Baseline Trust Matrix**: Initial trust scores are sampled for agents and tasks.
- **Attack Step**: Randomized noise is applied to produce the poisoned trust matrix.
- **Defense Step**: Partial restoration of trust through countermeasure logic.
- **QUBO Task Assignment**: Optimized agent-task assignment using NEAL sampling.
- **Results and Evaluation**: Comparison of trust rewards, assignment quality, and resilience metrics.

F. Implementation Details

The simulation framework was implemented in Python using NumPy, Matplotlib, Seaborn, and NetworkX for visualization, and Dimod for QUBO modeling and sampling.

Multiple simulation variants (v2, v3, v3.5) were created to test different scenarios:

- Number of agents and tasks (50 vs. 100 agents).
- Number of rounds (100, 500, 1000).
- Initial trust distributions and noise parameters.
- Defense restoration thresholds and penalties.

All versions share a consistent core pipeline and QUBO formulation.

G. Visualization Outputs

Simulation outputs include:

- Heatmaps of baseline, poisoned, and defended trust matrices.
- Bipartite assignment graphs overlaying QUBO solutions under attack vs. defense conditions.
- Trust score distributions after simulation rounds.
- Time-series plots showing the evolution of average fleet trust and individual agent trust.
- Bar charts comparing task success and failure counts over time.

These visualizations interpret the effects of attacks degrading system-wide trust, how defenses partially restore reliability, and how adaptive assignment improves coordination under adversarial conditions.

IV. Experimental Results and Evaluation

These simulated post-disaster scenarios evaluate the impacts of adversarial attacks, defense strategies, and quantum-optimized task assignments on trust-aware coordination. Each experiment focuses on a hospital rescue setting with multiple dynamic agents coordinating tasks under cyber-physical threat models such as GPS spoofing or communication tampering. Simulations are run with three focuses: v2 on attack and defense, v3 with hospital mapping, and v3.5 with a hybrid fleet-scale focus.

A. Trust Matrix Heatmaps: Baseline, Attack, and Defense (V2)

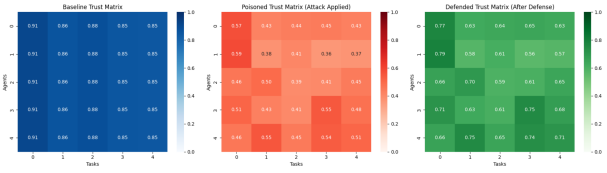

Fig. 2. Baseline Trust Matrix Heatmap (V2). Higher values indicate strong agent-task confidence pre-attack.

The baseline trust matrix in V2 represents the system's initial confidence in agent-task pairings before any attack occurs. This models a clean post-earthquake hospital coordination scenario where robots maintain reliable assessments of which units (supply bots, med-drones) are best suited to specific tasks. High trust scores here ensure optimal task allocation with minimal conflict.

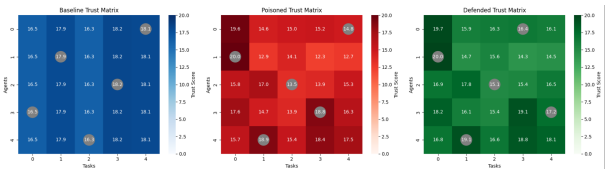

Fig. 3. Poisoned vs. Defended Trust Matrix Heatmap (V2) after attack and defense steps.

By simulating cyber attacks (e.g., GPS spoofing), trust is significantly degraded, as seen by lowered heatmap intensities. The "poisoned" matrix leads to suboptimal assignments that risk mission failure. Defense countermeasures partially restore trust by increasing low values where agent performance data indicates reliability. By modeling real-world cyber defense workflows, anomaly detection supports partial trust restoration in a post-disaster coordination setting.

B. Assignment Matrix Visualization with QUBO Optimization (V2)

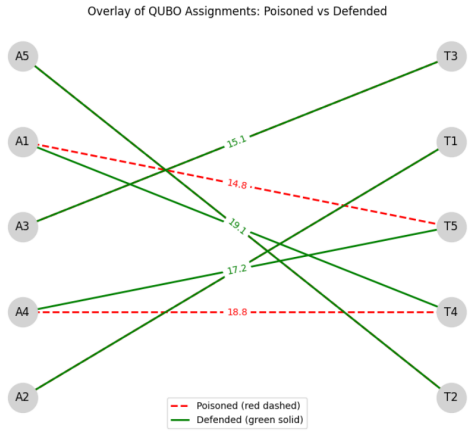

Fig. 4. Overlay of QUBO Assignment Decisions on Poisoned vs. Defended Trust Matrices (V2). Grey markers indicate selected assignments.

This overlay shows NEAL-sampled QUBO solutions on trust heatmaps, with grey circles marking the agent-task pairings selected after optimization. In the poisoned scenario, assignments are forced to optimize within degraded regions, reflecting constrained options with low confidence. After defense, the spread improves, allowing better allocation across partially restored trust regions. This highlights how quantum-inspired optimization can partially compensate for degraded trust under attack, modeling real hospital disaster-response workflows that constantly reassign critical delivery routes despite cyber threats.

C. Fleet-Wide Simulation Outcomes (V3.5)

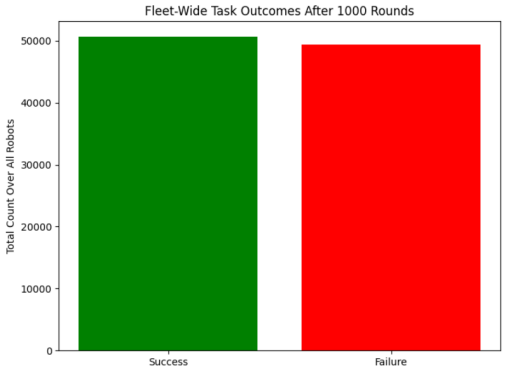

Fig. 5. Fleet-wide Task Outcomes After 1000 Rounds (V3.5): Success vs. Failure counts.

The success and failure bar chart shows task attempts over extensive simulation rounds in a hybrid scenario (V3.5). The experiment models a hospital-wide fleet of robots performing repeated deliveries, inspections, and rescues under irregular attacks. The relative success rate provides a quantitative evaluation of the system's resilience, showing that countermeasure-enabled coordination still retains significant task throughput despite adversarial attacks.

D. Trust Dynamics Over Time (V3.5)

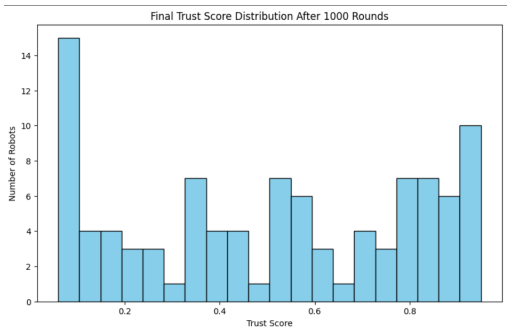

Fig. 6. Trust Distribution Histogram After Simulation (V3.5).

The spread of final trust scores among the robotic fleet shows most agents clustering at moderate to high trust levels, suggesting overall effectiveness of defense countermeasures. However, the visible tail toward low trust indicates a fraction of compromised or unreliable units that persist, reflecting the realistic need to quarantine or reassign compromised hospital robots post-attack.

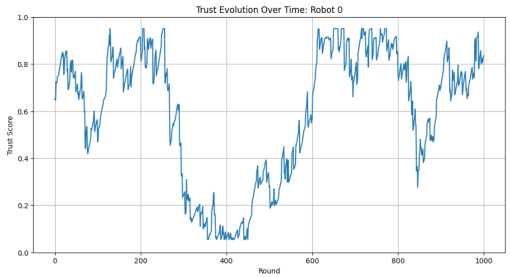

Fig. 7. Trust Evolution Over Time for a Single Robot (V3.5).

This time-series plot tracks trust evolution for one representative agent. The trust initially fluctuates but then stabilizes under defense countermeasures, validating the model's dynamic learning mechanism. It shows that trust is penalized after failures and gradually recovers with repeated success. In a post-disaster hospital, this reflects a single delivery robot learning to rebuild confidence despite network disruptions.

E. Average Fleet Trust Dynamics (V3.5)

This shows systemic trust trends across 1000 rounds, with general stabilization near mid-to-high trust levels, reflecting defense effectiveness. In a disaster setting, this would translate to sustained mission capability for the majority of robots despite periodic attacks.

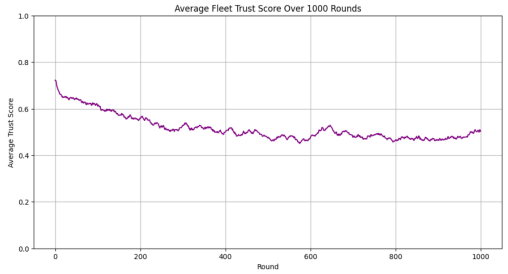

Fig. 8. Average Fleet Trust Over Simulation Rounds (V3.5).

F. Mission Outcomes and Compromised Robots Over Time (V3.5)

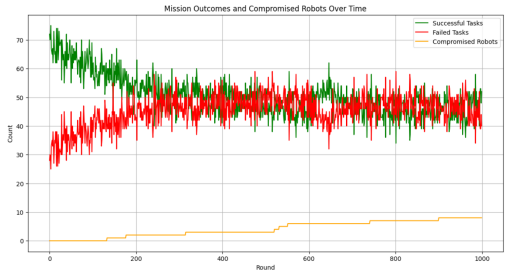

Fig. 9. Mission Outcomes and Compromised Robots Over Time (V3.5).

trace quantifies success, failure, and compromise counts for each round. The peaks in compromise rates highlight periods of intensified adversarial impact, while defense reduces their duration and effect on success rates. For hospital disaster-response collaboration, it suggests the system could provide early-warning indicators of attack waves and trigger adaptive rerouting to maintain delivery of critical supplies.

G. Discussion of Results Across Versions

Each simulation version explored different aspects of the problem:

- **V2** focuses on clear attack and defense parameters in trust matrices and QUBO assignments at a small scale, emphasizing the impact of poisoning attacks and defense countermeasures.
- **V3** emphasizes realistic 3D hospital mapping and task assignment, prioritizing environmental complexity.
- **V3.5** concentrates on larger fleet sizes and more rounds to model long-term fleet coordination, fleet-wide trust dynamics, and mission throughput under repeated attacks.

These results validate the core focus that trust-aware anomaly detection and defense encourage adaptive task allocation even under adversarial conditions. The simulations show degraded trust directly affects assignments, while partial restoration plus QUBO optimization recover mission effectiveness. This models a post-disaster hospital fleet that can keep critical services operational even in the face of cyber-physical disruptions.

V. Conclusion

...

VI. Future Work

...

Acknowledgments

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.

References

[1] A. Bojchevski and S. Günnemann, "Adversarial Attacks on Node Embeddings via Graph Poisoning," arXiv preprint arXiv:1809.01093, 2018. [Online]. Available: https://arxiv.org/abs/1809.01093

[2] A. Haskard and D. Herath, "Secure Robotics: Navigating Challenges at the Nexus of Safety, Trust, and Cybersecurity in Cyber-Physical Systems," ACM Computing Surveys, Mar. 2025. [Online]. Available: https://doi.org/10.1145/3723050