Problem Statement

As robots become increasingly autonomous and are deployed in complex, real-world environments, their ability to adapt to novel, unforeseen situations becomes critical. Traditional robotics systems rely heavily on hard-coded rules or pre-defined paths, which are insufficient for dynamic and unstructured environments. Recent advancements in artificial intelligence—particularly reinforcement learning (RL) and imitation learning (IL)—enable robots to learn from interactions and adjust their behaviors without explicit programming. This shift toward adaptive behavior significantly enhances functionality but also introduces new security vulnerabilities and trust concerns.

Autonomous robots must process large volumes of sensor data, make real-time decisions, and interact with the physical world. If these systems are compromised—through sensor spoofing, adversarial attacks on learning models, or unauthorized command injections—the consequences can be severe, affecting both safety and performance. Furthermore, the black-box nature of many learning-based systems undermines user trust and raises questions about transparency, accountability, and resilience under attack.

This research addresses the urgent need to secure dynamic robotic behavior by developing a dual framework that:

1. Enables robots to learn and adapt to unexpected events and environments, and
2. Embeds robust cybersecurity mechanisms to defend against manipulation, ensuring the system's integrity, reliability, and trustworthiness.

Literature Review 1

Title: Secure Robotics: Navigating Challenges at the Nexus of Safety, Trust, and Cybersecurity in Cyber-Physical Systems

Authors: Adam Haskard & Damith Herath

Publication Year: 2025

Key Ideas

The article introduces secure robotics as a unified interdisciplinary framework that integrates trust, safety, and cybersecurity, domains that have traditionally been analyzed in isolation. This approach responds to the growing presence of cyber-physical threats and the increasing need for robotic systems that are resilient, adaptable, and trustworthy. As robots are increasingly deployed in real-world environments, ensuring human safety and adhering to ethical principles become crucial. The authors reference Asimov’s laws to stress that robotic systems should align with human values and ethics. The article discusses the importance of integrating societal norms into robotic systems to facilitate smoother interaction within human environments. Trust is presented not as a binary concept but as something that evolves over time. Human-robot interaction (HRI) is shaped by psychological models of trust and risk management, with an emphasis on understanding the contextual factors that affect trust development. In augmented reality systems, where virtual elements are superimposed onto the physical world, unique threats arise, such as sensory spoofing, unauthorized access, and cyberattacks. The authors argue that combining cyber-physical systems (CPS) with the computational capabilities of information systems (IS) enables better interaction with the physical world. A taxonomy of trust-relevant failures in HRI is provided, including design failures, system failures, expectation mismatches, and user errors, each influencing trust differently.

Contributions

This article contributes by defining secure robotics as a new interdisciplinary field that merges trust, safety, and cybersecurity within CPS. It outlines the cybersecurity challenges and vulnerabilities that robots face, especially when trust and safety intersect. Challenges such as weak data integrity, insecure networks, and ineffective human-machine collaboration are linked to poor communication and weak authentication. To enhance system reliability, the authors suggest improving transparency, clearly communicating intent, and increasing dependability. The article highlights threats to human-robot trust caused by cyberattacks like denial-of-service, dynamic manipulation, and replay attacks. It proposes a conceptual framework to understand and mitigate trust failures in robotic systems, particularly those involving unauthorized access, data manipulation, or risks of physical harm.

Limitations

Since this unified framework is relatively new, it remains theoretical and lacks experimental validation. The article provides conceptual models but does not offer hands-on implementations or real-world testing. While it raises cybersecurity concerns, it does not go into detail about specific defense tools or system-level implementations. Furthermore, the article does not provide concrete guidelines or alternatives for addressing ethical concerns in robotic system development.

Extend/Improvement

To build on this article, my research will focus on implementing security-aware learning systems for adaptive robots operating in unpredictable environments. By integrating trust and cybersecurity during the learning process, I aim to reinforce learning pipelines, such as imitation or reinforcement learning, where trustworthiness is treated as a core goal. Drawing on the taxonomy of failures presented in the article, I plan to design an adaptive feedback system that evaluates and addresses communication gaps affecting user trust during the robot's learning phase.

Literature Review 2

Title: Artificial Intelligence and Machine Learning Enhance Robot Decision-making Adaptability and Learning Capabilities Across Various Domains

Authors: Md Delwar Hussain, Md Hamidur Rahman & Nur Mohammad Ali

Publication Year: 2024

Key Ideas

This article presents the fundamentals of how AI and ML enable robots to perceive, learn, and reason in real time. Using artificial neural networks, robots can extract information from sensor data, recognize patterns, and make predictions, which allows them to function in unpredictable environments. By adapting their behavior based on data and context, robots are tuned for use in domains such as industrial automation, healthcare, space exploration, and disaster response. The article emphasizes three types of learning: supervised learning (with labeled datasets), unsupervised learning (identifying patterns in unlabeled data), and reinforcement learning (where agents learn from trial and error). It reinforces the idea that with real-time sensory data, AI- and ML-powered robots can dynamically adjust their behavior in response to environmental changes. Case studies like the Da Vinci Surgical System and Boston Dynamics' Spot Robot demonstrate the potential of robots to handle complex tasks.

Contributions

The article provides a broad overview of how AI and ML enhance robotic decision-making across multiple complex domains. It shows that with learning models, AI-powered robots can respond adaptively to real-world settings such as healthcare or disaster environments. It also highlights how machine learning enables robots to learn both from pre-collected offline data and through real-time interaction with their environment.

Limitations

While informative, the article is survey-based and does not delve deeply into technical implementations. Although it briefly mentions cybersecurity concerns and trust issues, these are not central themes. The case studies presented highlight high performance but lack transparency about system failures, adversarial risks, or real-world operational limits.

Extend/Improvement

My focus will expand on the topics of security and trustworthiness in robotics using reinforcement learning (RL) to ensure reliable operation in unpredictable environments. I also plan to explore safety mechanisms against failures by examining adversarial examples and their effects on learning systems. Additionally, I aim to bridge the gap between cybersecurity and trust, an area the article does not deeply address.

Literature Review 3

Title: Adversarial Attacks on Neural Network Policies

Authors: Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, & Pieter Abbeel

Publication Year: 2017

Key Ideas

The article focuses on deep reinforcement learning (DRL), where neural network policies are vulnerable to adversarial attacks, just like in supervised learning. It shows how small perturbations can cause a robot to make incorrect decisions, even if it was previously well-trained. The paper outlines two types of attacks, white-box attacks, where the attacker has full knowledge of the policy network and can create targeted perturbations, and black-box attacks, where the attacker has no internal knowledge but exploits transferability to generate adversarial examples that still mislead the policy.

Contributions

This article systematically demonstrates how adversarial examples can significantly affect neural network policies trained with reinforcement learning. It shows that even agents trained in controlled environments like games can fail due to minor input changes. The paper introduces the Fast Gradient Sign Method (FGSM) as a way to create adversarial observations both during training and test time. It also provides experimental analysis across different RL algorithms (DQN, A3C, TRPO) and multiple environments to show how performance drops under adversarial conditions.

Limitations

While the results are compelling, the experiments were only conducted in simulated environments, not in real-world settings. The study focuses on short-term technical performance and doesn't explore human-robot trust implications. It highlights vulnerabilities but doesn't provide defensive strategies. The attacks also only apply to visual input policies, without addressing more complex sensor fusion scenarios common in real-world robotics.

Extend/Improvement

My work will focus on building defensive strategies for adaptive robots in unpredictable, real-world environments, beyond simulations. I plan to develop methods for detecting and mitigating adversarial inputs during runtime for both white-box and black-box attacks. Since the article emphasizes the vulnerability of neural policies, I will explore learning pipelines that enable robots to adapt to these attacks and maintain reliable performance.

Planned Weeks

Week 3 (June 10–14)

- Refine RL/IL training environment in Colab
- Implement basic adversarial attack simulations
- Document results and challenges in GitHub
- Begin exploring trust scoring logic

Week 4 (June 17–21)

- Build out initial trust feedback
- Test learning model under adversarial input conditions
- Compare behavior with/without defenses
- Upload progress in GitHub

Week 5 (June 24–28)

- Integrate basic transparency elements
- Continue code activity

Week 6 (July 1–5)

- Stress-test system: multi-scenario simulations
- Continue working on final presentation and report

Week 7 (July 8–12)

- Finalize all experimental runs and collect data
- Test code and results

Week 8 (July 15–19)

- Revise final report based on feedback
- Finalize presentation slides
- Clean and organize GitHub repository

Week 9 (July 22–26)

- Practice final presentation
- Clean and organize GitHub repository

Final Week (July 29–31)

- Final presentation
- Program wrap-up

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.