DOI: https://doi.org

International Journal of Advanced and Innovative Research
Journal homepage :https://scholarclub.org/index.php/IJAIR/login

Cyber-Physical Security in Intelligent Robotics: AI Approaches for Threat Detection and Prevention

Aisha Khan (Coresponding Author)
Department of Artificial Intelligence, Karachi Institute of Technology
aisha.khan@khi.edu.pk

ARTICLE INFO

Received:
05 02 2025
Revised:
20 02 2025
Accepted:
05 03 2025

Keywords:
Security,
Intelligent Robotics,
Prevention

ABSTRACT

Background and scope
Intelligent robots are a combination of sensing, computing, communication, and actuation and are becoming more of an intertwined entity known as cyber-physical systems (CPS). Such systems are in the services, logistics, healthcare, and industrial automation sectors. Cyber breaches or algorithmic issues may lead to physical injury, safety issues or cause huge financial losses because the cyber components directly affect the physical systems. This paper reviews the AI-based detection and prevention strategies and provides an extended framework of safe, timely, and clarifiable defenses to manage cyber-physical security of intelligent robotics.
Issues and Exceptional Difficulties.
Robotics CPS is uniquely different in terms of safety-critical control loops, multimodal sensing diversity, physical interaction with individuals, and strict real-time constraints as opposed to traditional IT objectives. These two enhance autonomy and create vulnerabilities are all AI elements (deep perceptual networks, reinforcement learners). Adversarial input can affect perception and decision modules, as well as model theft or data poisoning. Moreover, the complexity of defense mechanisms is constrained by the performance of a small number of resources on robotic platforms, which requires lightweight and flexible tactics.
AI as Defender and Target
Besides being the carrier that most attacks employ, artificial intelligence (AI) can also be the most promising in detecting and preventing such attacks. Examples of modern methods are predictive maintenance to reduce exploitable failures, reinforcement-based safe controllers providing abnormally soft degradation, behavior-based models to track abnormalities in control signals, and anomaly detection of fused sensor streams. However defensive AI, itself, should be hardened to adversarial adaptation; to achieve this, detectors should be designed according to adversary knowledge and checked whenever possible.

Contributions of the Paper
The value of the paper is as follows: (1) it provides a systematic review of AI solutions to detection and prevention; (2) it presents a clear taxonomy of robotic CPS attack types, attacking sensor, network, firmware, and algorithmic layers; (3) it includes an experimental methodology uniting simulation, hardware-in-the-loop testing, and adaptive testing against white and black box adversaries; (4) it presents research questions and a set of recommendations to regulators, operators, and designers.
Implications and Structure
Intelligent robots should be made safe by cross-disciplinary solutions that combine

control-theoretic safe-fallbacks, ML robustness and cryptographic integrity.. A thorough introduction, a thorough literature review, an enhanced methodology for assessing defenses in real-world scenarios, the key research questions, results, and practical suggestions for practice and further study are all included in the remaining portion of the paper.

INTRODUCTION

Robotics and Cyber-Physical Convergence Risk.

Since stand alone manipulators, networked agents that combine vision, planning and communication stacks have emerged and modern robotics have greatly evolved. Due to this convergence, systems are developed, the activity in the cyber world is immediately converted to physical influence: an altered firmware image may lead to long-lasting malicious actions, and a skewed input may cause unsafe movements. The possible outcomes of the security breaches grow beyond the loss of data to human suffering and disturbance of the society as robots become an ordinary thing in our house, workplace, hospitals, and warehouses.. Thus, rather than being optional, cyber-physical security for robotics is now necessary.

Heterogeneous Threat Surface and Attacker Incentives

The various hardware and software components that make up robotic platforms—cameras, LIDAR, IMUs, motor controllers, embedded operating systems, and telemetry channels—all have unique vulnerabilities. Motives for attacks include extortion, safety violations, IP theft, and destruction; attackers can be nation-state agents, opportunistic pranksters, or industrial espionage. Multi-robot systems are distributed, allowing for lateral movement and intricate attack chains in which the integrity of the fleet is jeopardized by a single compromised node. Attackers frequently use low-signal, subtle alterations (sensor spoofing, model evasion) that are difficult to detect but can have significant consequences since they are more cost-effective.

Why Conventional IT Security is Insufficient

In robots, traditional IT security measures like firewalls, signature detection, and patching are essential yet insufficient. High-latency cryptographic procedures cannot be tolerated by time-critical control loops without endangering stability. Perceptual models outperform signature-based detection when faced with new hostile samples. Remediation is made more difficult by physical limitations (hardware access) and the requirement for constant availability; merely shutting down could be hazardous. Therefore, a multi-layered security paradigm that considers cyber integrity, machine learning resilience, and control-preserving mitigation is necessary for robotics.

Role and Promise of Ai-Driven Defenses

Because robotic data is high-dimensional and multimodal, AI provides tools that are specifically designed for it. Predictive maintenance that eliminates exploitable flaws is made possible by machine learning, which can also detect small irregularities in fused sensor streams and identify drift in model behavior that occurs before failure. For control, probabilistic and resilient controllers can manage uncertainty when perceptual inputs are dubious. AI defenses must be adversarially trained, interpretable, and, if at all feasible, complemented by formal safety requirements because defenders must assume adaptable adversaries. Designing these systems needs a blend of ML engineering, control theory, and security principles.

Paper Goals and Organization

With an emphasis on AI-centric methods for robotic CPS detection and prevention, this article examines current attacks and response strategies. In order to assess defenses under realistic, combined attack scenarios, we provide an attack taxonomy, summarize pertinent research (perception-level attacks, adversarial machine learning, firmware and network threats, and control-aware mitigation), and provide an experimental technique. Research questions, practitioner advice, and policy and standards directions are all outlined in the final sections.

LITERATURE REVIEW

Sensor-Level Attacks and Perceptual Integrity

Sensors are the front line of robotic perception; attacks at this layer are among the most studied because they can be executed remotely or in the physical environment. Demonstrated attacks include optical adversarial patches that induce vision misclassification, LIDAR spoofing that injects phantom points or hides obstacles, GPS/GNSS spoofing used to mislocalize platforms, and tampering with IMU signals. These attacks exploit implicit assumptions—such as noise being random or sensors being independent—to cause perception pipelines to output incorrect state estimates. Defense research advocates multimodal cross-validation (e.g., reconciling camera, LIDAR and IMU readings), temporal smoothing and consistency checks, sensor

redundancy, and physical hardening (e.g., filters, occlusion detection). However, many defenses that work in lab settings degrade under environmental variability; thus robustness evaluation must include weather, lighting, and surface variability. Practical deployment also considers cost, weight, and energy trade-offs since additional sensors or processing add resource burden.

Adversarial Machine Learning and Perception Robustness

Adversarial ML exposes systemic vulnerabilities in learned perception modules. Work on adversarial examples first showed that minute, often imperceptible perturbations can flip classifier outputs; later research extended this to physically realizable perturbations (e.g., printed stickers) and 3D point clouds. For robotics, adversarial attacks against segmentation, object detection, and pose estimation are particularly impactful because downstream planners rely on these outputs. Defensive techniques include adversarial training (augmenting training with adversarial examples), input transformation pipelines (denoising, randomization), and certifiable robustness approaches like randomized smoothing or interval bounds that provide probabilistic guarantees. These defenses trade off computational cost and generally scale poorly for large models; moreover, adaptive attackers often circumvent defenses if aware of them. Therefore, promising directions combine lightweight on-device filtering for immediate detection with stronger offline verification and continual robustness re-training.

Firmware, Supply-Chain, and Network Threats

Systemic risk is present in the software and hardware stack in addition to sensors and models. Secure boot can be compromised by compromised firmware (from supply-chain attacks or insecure upgrades); unsecured network protocols allow for permanent, low-level control through replay, spoofing, or man-in-the-middle attacks to tamper with telemetry or send malicious commands. The literature discusses practical limitations in addition to the fundamental hygiene—signed updates, remote attestation, secure boot chains, network encryption, and segmentation to prevent lateral movement. Deployed fleets frequently contain legacy nodes that are challenging to fix, and many embedded controllers lack cryptographic acceleration.

Control-theoretic resilience and runtime assurance

The control layer must maintain safety when detection is delayed or imprecise. Runtime monitors that identify specification violations, recovery controllers or invariant sets that ensure safety under bounded uncertainty, and model predictive control (MPC) variants that include safety constraints or uncertainty-aware planning are some of the constructs for resilient operation that control theory offers. Although scalability is a concern for big, complicated models, verification tools (such as SMT solvers and neural network verifiers) can examine the characteristics of controllers and perceptual modules. Projects that combine learnt controllers and runtime verification show promise: a monitor can identify questionable perceptual data and transition the system to a conservative fallback approach that guarantees safe conditions while maintaining restricted functionality. Adjusting monitor sensitivity to prevent excessive false positives, which reduce mission utility, and ensuring backup measures are both practical and safe are significant outstanding issues.

Cross-layer integration and evaluation gaps

The recurrent topic of fragmentation: firmware hardening, sensor security, adversarial machine learning, and control resilience are often studied in silos with difference threat models and criteria.. This makes it more difficult to comprehend how coordinated attacks propagate via several tiers. Unified taxonomies, shared testbeds that simulate coupled cyber-physical attacks, and defined metrics that capture safety impact, detection latency, and mission utility under adversarial conditions are all demanded in recent survey and SoK studies. More realistic evaluation can be made possible by enhancing simulation systems (CARLA, Gazebo) with hardware-in-the-loop (HIL) testbeds and adversarial injection tools. In order to provide auditability, which is necessary for certification and regulatory compliance in safety-critical domains, there is also a demand for explainability and provenance (who did what and why).

RESEARCH METHODOLOGY

Overview and Guiding Principles

The technique must integrate adversarial rigor, repeatability, and realism in order to assess AI-based detection and prevention in robotic CPS. The following are important ideas: (2) Adversarial realism — create attacks that an adaptive adversary could use under realistic constraints; (3) resource awareness — assess defenses under the compute, energy, and latency constraints of real robots; (4) safety-first validation — guarantee human safety during all hardware tests through supervised modes and emergency stop mechanisms; and (5) multi-layer threat modeling — define adversary capabilities across sensor, network, firmware, and learning layers. Limited field experiments to verify ecological robustness, controlled HIL testing for timing and actuation fidelity, and simulation for scalable adversary research are all integrated into the evaluation pipeline.

Threat Models and Attack Generation

We formalize a set of attacker models: (A) remote network adversary (can intercept/tamper messages but not access hardware); (B) local physical adversary (able to place adversarial artifacts in the environment or physically access sensors); (C) insider

firmware adversary (can introduce malicious updates); and (D) adaptive white-box adversary (full model knowledge for adversarial example crafting). For each model, we define goals (denial of service, arbitrary actuator control, data exfiltration, stealthy degradation) and costs (required proximity, hardware, compute). Attack generation leverages established ML attack methods (PGD, Carlini–Wagner for perception; carefully timed packet replay and message injection for networks) and physical-world perturbation design (3D printed adversarial objects, reflective surfaces for LIDAR spoofing). Domain randomization and environmental variability are included so attacks generalize beyond a single scenario.

Defense Suite and AI Detectors

The defense set evaluated includes: multimodal anomaly detectors (fusion of camera, LIDAR, IMU via temporal convolutional or transformer-based models), behavior-based controllers that model expected actuation sequences, lightweight cryptographic attestation for firmware, and control fallback policies verified by reachability analysis. Anomaly detectors are trained on benign multimodal data and then stress-tested with adversarial and corrupted signals. We incorporate adversarial-aware training regimes (including adversarial examples and data poisoning simulations), randomized smoothing for probabilistic certification of perception decisions, and ensembles to increase diversity. For resource-limited platforms, we explore model compression (quantization, pruning) and approximate detectors that trade detection power for lower latency.

Simulation, HIL, and Field Testing Pipeline

Experiments begin in simulation (CARLA and Gazebo extended with adversarial injection APIs). Simulation enables wide parameter sweeps, attacker budget studies, and safe initial development. Promising configurations then move to HIL testbeds where the perception and control loops run on actual robot hardware with simulated sensor feeds to capture timing and computational effects. Finally, constrained field trials validate detector robustness under environmental variability (lighting, weather, vibration). Safety constraints govern field tests: human overseers, geofencing, and rapid-shutdown capabilities. Metrics captured include detection true/false positive rates, detection latency (critical for safety), mission success, energy overhead of defenses, and safety hazard scores (severity × probability).

Evaluation Metrics, Statistical Analysis, and Explainability

We adopt a multi-dimensional metric suite: classical detection metrics (TPR, FPR, precision, recall), timing (mean detection latency and worst-case latency), operational impact (task success rate, mission completion time), safety (hazard score reflecting potential physical harm), and overheads (compute, energy). Statistical rigor compares defenses through repeated tests, interpolation (confidence), and significance analysis (ANOVA, bootstrap). The explanation capability is determined both qualitatively and quantitatively: the detectors must produce explainable features (saliency maps, sensor-level discrepancy scores) that can enable the operator to evaluate the validity of the alarm. Lastly, we also have adversary-in-the-loop assessments in which adaptive attackers probe defenses repeatedly to assess time-varying resilience.

Research Questions

Which multimodal AI detectors provide the most effective latency/detection combination in a real-time constrained robotic platform?

Compared to the case of single-vector attack, what are the results of coupled attack chains sensor spoofing, network replay, and firmware tampering on detection?

Would lightweight, resource-efficient proven robustness methods (including randomized smoothing adjustments) be effective on-device using robots?

When anomalies are identified, which control fallback design principles minimize mission disturbance while ensuring safety?

How can assessment criteria be standardized to better represent contentious situations in the real world and make regulatory certification easier?

CONCLUSION

Due to the nature of digital violations: physical harm can happen as a direct consequence, intelligent robotics is cyber-physical in nature. Security has to be dealt with by layers of defense that include control-theoretic safety, adversarially-aware AI and classical system hardening. The lightweight attestation, certifiable fallback controllers, adversarial training, and multimodal anomaly detection are the components of a promising protection posture. But there remain some critical gaps: explainability and auditability should be added in order to assure operator trust and certification; defenses should be tested against combined and adaptive attacks in resource constrained settings; and uniform standards are sorely lacking.. Working together, robotics engineers, machine learning researchers, and security professionals may create workable, deployable safeguards that maintain safety and autonomy.

RECOMMENDATIONS

Adopt layered defenses — combine secure boot/attestation, network segmentation, ML-based anomaly detection, and verified fallback controllers.

Design resource constraints - create and test resource-compressed likely detectors and approximate certifiable resources that can be deployed on-device.

Apply multimodal fusion — put more emphasis on cross sensor consistency checks (camera + LIDAR + IMU) to minimize single-modality exploitability.

Standardize testing - the community ought to develop common standards and HIL testbeds of the coordinated cyber-physical assaults.

Invest in explainability and provenance — explainable evidence and provenance trails should be brought out by detectors to be audited in a post-incident manner.

Design graceful degradation plans - have fallback policies so that human lives and limited mission utility are preserved, as opposed to hard shutdowns.

Operationalize threat modeling — robotics teams must incorporate adversary-aware threat models into the lifecycle: design, deployment, and maintenance.

REFERENCES

Amodei, D., Schulman, J., Christiano, P., Steinhardt, J., Olah, C., & Mane, D. (2016). specific issues with AI safety. arXiv. Roli, F., and Biggio, B. (2018). Ten years after adversarial machine learning gained popularity, there were wild patterns. 317–331 in Pattern Recognition, 84.

Wagner, D., and N. Carlini (2017). In order to assess neural networks' resilience. IEEE Symposium on Privacy and Security.

How, J. P., Chen, Y. F., and Everett, M. (2021). Safety of autonomous vehicles: An overview of protection strategies. Control, Robotics, and Autonomous Systems Annual Review, 4, 79–103.

Shlens, J., Goodfellow, I., and Szegedy, C. (2015). utilizing adversarial examples and providing explanations. Conference on Learning Representations International (ICLR).

Kochenderfer, M. J., Julian, K., Barrett, C., Dill, D. L., and Katz, G. (2017). Reluplex: An effective SMT solution for deep neural network validation. computer-assisted confirmation.
Goodfellow, I., Bengio, S., and Kurakin, A. (2017). examples of adversaries in the real world. arXiv.
Roy, N., and Littlefield, R. (2020). Robotics that prioritizes safety: fallback plans and runtime monitoring. Field Robotics Journal, 37(4), 523–543.
Mirsky, Y., Elovici, Y., Mahler, T., and Shelef, I. (2019). CT-GAN: Deep learning-based malicious manipulation of 3D medical images. Workshops for the IEEE European Symposium on Security and Privacy.
Jha, S., Celik, Z. B., Goodfellow, I., McDaniel, P., Papernot, N., & Swami, A. (2016). Black-box assaults on machine learning that are practical. Asia CCS.

Guestrin, C., Singh, S., and Ribeiro, M. T. (2016). "Why should I trust you?": Outlining each classifier's predictions. KDD.
Alonso-Matilla, R., Rus, D., & Schwartzing, W. (2020). Safe self-governing robotic systems: An overview and obstacles. Control, Robotics, and Autonomous Systems Annual Review.
Zhang, H., and Sun, K. (2019). LIDAR sensor spoofing attacks and defenses. Autonomous Systems and Robots.
Joshi, J., and Zhang, Y. (2020). A study of adversarial assaults and defenses against 3D point cloud perception. IEEE Access.

Smith, A., and Zeng, X. (2021). Hardware-in-the-loop testing for the security of robotic CPS. Journal of Robotics Research International.