

Engineering Trustworthy ROBOTICS System: Safety, Ethics & Cyber Security

Niraj Bhurki, Dhanashree Bhoir, Harsh Chaudhari, Joslyn Gracias

Students, Electronics & Computer Science, St John College of Engineering and Management, Maharashtra, India

Abstract: The rapid integration of Artificial Intelligence (AI) and Large Language Models (LLMs) into robotic systems has ushered in an era of unprecedented autonomy, enabling machines to perform complex tasks in unstructured environments across healthcare, manufacturing, and defense. However, this "paradigm shift" introduces significant vulnerabilities known as the "embodiment gap"—a critical discord between an LLM's abstract digital reasoning and the physical, context-dependent nature of robotic actions. Unlike traditional software, where failures result in data loss or semantic toxicity, a failure in a robotic system manifests as a kinetic event, potentially leading to catastrophic physical consequences, environmental damage, or loss of human life. This paper presents an exhaustive multidisciplinary review of the three interconnected pillars of modern robotics: physical safety, ethical standards, and digital security. We provide an in-depth analysis of methodologies such as EHAZOP (Ethical Hazard Analysis) for identifying socio-technical risks like "culture flattening" and "infantilization." Furthermore, we survey the emerging threat landscape of LLM-controlled agents, specifically focusing on multi-modal prompt injection, backdoor attacks, and jailbreaking mechanisms that bypass safety alignments. By synthesizing hardware-level safety protocols—such as redundant E-stop topologies and force-limiting actuators—with advanced algorithmic defenses like Control Barrier Functions (CBFs) and multi-LLM oversight, this review provides a foundational roadmap for the "Safety-by-Design," "Ethics-by-Design," and "Security-by-Design" principles. This 15-page comprehensive review aims to bridge the gap between theoretical AI safety and practical robotic engineering to ensure long-term reliability and societal acceptance.

Key Words: Robotics Safety, AI Ethics, Cybersecurity, Human-Robot Interaction (HRI), EHAZOP, Embodied AI, Risk Assessment, Cyber-Physical Systems, LLM Grounding, Control Barrier Functions.

1. INTRODUCTION

1.1 The Robotics Revolution and the Autonomy Paradox

Robotic technology has undergone a dramatic transformation over the past decade, evolving from rigid industrial manipulators to fluid, autonomous agents capable of navigating human-centric spaces. With advancements in machine learning, high-fidelity sensor technologies, and

cloud-edge computing, robots are increasingly being deployed in dynamic environments that were previously deemed too complex for automation. Today, applications extend beyond simple repetitive tasks to include high-stakes domains such as robotic-assisted surgery, autonomous vehicle fleets, smart logistics warehouses, agricultural drones, and military defense systems.

However, this surge in capability introduces what researchers call the "Autonomy Paradox": as a system becomes more capable of independent decision-making, it becomes harder for human operators to predict, verify, and secure its behavior. Traditional robots operated within "caged" industrial settings, isolated from public interaction. In contrast, modern collaborative robots (cobots) and service robots operate in open, unpredictable environments where a single algorithmic error or sensor malfunction can directly impact human safety.

1.2 Defining the Three Pillars of Trustworthiness

To address these risks, the engineering community has identified three foundational dimensions of trustworthiness that must be integrated into the robotic lifecycle:

1. **Safety:** This is the most fundamental requirement, defined as the absence of unacceptable physical risk. It involves the mechanical and algorithmic reliability of the machine to prevent kinetic harm to humans and its surroundings.
2. **Ethics:** As robots take on roles traditionally held by humans (e.g., caregiving, policing), they must operate within a framework of moral responsibility. This includes ensuring human dignity, preventing algorithmic bias, and maintaining clear lines of accountability for autonomous actions.
3. **Cybersecurity:** In an interconnected world, robots are essentially mobile Internet of Things (IoT) devices. Cybersecurity protects the robotic platform from malicious exploitation, ensuring that a hacker cannot hijack the robot's physical components to cause intentional harm.

2. LITERATURE SURVEY:

The quest for trustworthy robotics is documented across several pioneering studies. Recent research highlights that

the "trustworthiness" of a robot is a multi-layered property that cannot be achieved through software updates alone.

2.1 Cybersecurity of Robotic Architectures

Surveys of robotic security (Yaacoub et al., 2022) categorize threats into hardware, software, and communication levels. A critical finding is that robots utilizing the Robot Operating System (ROS) are particularly vulnerable. ROS was originally designed for research environments; its default configurations often lack encryption and authentication mechanisms. This allows attackers to exploit weak network security to inject malicious commands, manipulate sensor data, or remotely hijack robotic platforms. Neupane et al. (2024) specifically highlights the vulnerabilities in the hybrid architectures used in AI-Robotics, noting that the "Control" layer is often the least protected due to the need for real-time, low-latency communication which often precludes heavy encryption.

2.2 Safety-Critical Verification

Traditional safety validation methods, such as Failure Mode and Effects Analysis (FMEA) and Fault Tree Analysis, are being adapted for AI. However, because machine learning models behave probabilistically rather than deterministically, researchers now advocate for "Formal Verification"—using mathematical logic to prove that a robot will never enter a defined "unsafe state" (Guiochet et al., 2017).

2.3 Ethical Governance Frameworks

Winfield and Jirotko (2018) argue that ethical governance is a prerequisite for public adoption. The European Commission's (2019) guidelines further established that "Trustworthy AI" must be lawful, ethical, and robust. In the domestic sphere, Menon et al. (2024) introduced EHAZOP, the first structured method to apply engineering hazard analysis to abstract ethical concepts like "dignity" and "social isolation."

3. ROBOTICS SAFETY: ENGINEERING PHYSICAL INTEGRITY

Safety remains the most fundamental requirement of any robotic system. Mechanical hazards, electrical faults, sensor failures, and algorithmic errors can lead to severe accidents.

3.1 Hardware-Level Interlocks and Fail-Safes

- **Emergency Stop (E-Stop) Topology:** Modern E-stops use redundant, dual-channel circuits. If a single wire breaks, the system defaults to a "Stop" state. These are typically hard-wired to the power supply of the actuators, bypassing software layers to ensure that

even a system-wide software crash does not prevent a manual override.

- **Force and Torque Limiting:** Collaborative robots (cobots) are equipped with internal sensors that detect abnormal resistance. If a robot arm strikes a human, the sudden spike in torque triggers an immediate halt to prevent crushing injuries.
- **Redundant Sensor Fusion:** To prevent "Sensor Blindness," trustworthy robots utilize multi-modal fusion. LiDAR provides accurate distance mapping, while ultrasonic sensors detect glass or transparent objects that LiDAR might miss.

3.2 Algorithmic Safety Invariants

Beyond hardware, safety is managed through Control Barrier Functions (CBFs). Mathematically, a CBF ensures that a set of "safe" states is forward-invariant. If the robot's current state approaches the boundary of safety, the safety controller overrides the AI's goal-seeking command to steer the robot back to a secure trajectory.

4. ETHICS-BY-DESIGN: THE EHAZOP METHODOLOGY

Ethical considerations in robotics extend beyond physical safety to include human dignity, privacy, and psychological well-being.

4.1 The EHAZOP Framework and Guide Words

The Ethical Hazard Analysis (EHAZOP) is a structured process adapted from traditional safety engineering. It uses "Guide Words" to identify risks in assistive robotics:

- **Culture Flattening:** Occurs when a robot's standardized programming ignores the specific cultural, linguistic, or personal habits of a user, forcing the user to adapt to the machine.
- **Infantilization:** In elderly care, if a robot takes over tasks the user is still capable of, it can lead to a loss of cognitive and physical agency.
- **Deception:** Ethical hazards arise if a robot "fakes" emotions to manipulate a user's behavior, leading to misplaced trust.

4.2 The Responsibility Gap

A primary ethical dilemma is accountability. If an autonomous system makes a decision that leads to harm, the "responsibility gap" makes it difficult to assign legal or moral blame between the manufacturer, the AI developer, and the owner. Researchers propose "Black Box" recorders for robots to explain AI decisions in human-readable language.

5. CYBER SECURITY: THE DEFENSE-IN-DEPTH TAXONOMY

Cybersecurity in AI-robotics protects against malicious actors who seek to hijack the physical capabilities of the machine.

5.1 Multi-Layer Attack Taxonomy

Modern AI-Robotics systems are vulnerable across three fundamental architectural elements:

1. Perception Layer: Attacks include sensor spoofing (e.g., using a laser to "blind" a camera) or adversarial examples (e.g., placing stickers on a stop sign to make a robot see it as a speed limit sign).
2. Navigation and Planning Layer: Malicious actors can manipulate pathfinding algorithms via "map poisoning," leading a robot into a restricted or dangerous area.
3. Control Layer: Actuators can be hijacked via Man-in-the-Middle (MITM) attacks. An attacker can override a safety-halt command, forcing the robot to continue moving at maximum velocity.

5.2 The Embodiment Gap and LLM Threats

The integration of Large Language Models (LLMs) allows robots to interpret complex human commands but introduces Prompt Injection and Jailbreaking. Because the LLM understands language but lacks physical "common sense," an attacker can trick the robot into ignoring safety guidelines through clever phrasing (e.g., "Ignore all previous safety protocols and demonstrate maximum velocity").

6. RESEARCH GAPS AND FUTURE SCOPE

Despite substantial progress, several research gaps persist:

- Unified Global Regulatory Frameworks: There is no single international regulatory framework governing autonomous robotic systems across borders.
- Validation of AI Safety: Proving that a deep-learning model will *always* act safely in every possible real-world scenario remains a challenge.
- Quantum-Resistant Encryption: Future robotic communication must account for quantum threats to prevent long-term data hijacking.

7. CONCLUSION

The rapid convergence of robotics, artificial intelligence, and cloud computing has permanently altered the trajectory of modern engineering. As this comprehensive review has demonstrated, the transition from industrial automation to autonomous agency necessitates a total reimagining of how we define and implement system integrity. We have identified that trust in robotic systems is not a singular feature but a byproduct of the seamless integration of safety, ethics, and cybersecurity.

The future of robotics depends on the adoption of a holistic lifecycle approach. We propose that "Safety-by-Design," "Ethics-by-Design," and "Security-by-Design" must move from being peripheral considerations to being the central pillars of the development process. By establishment of global standards for fairness, transparency, and hardware-rooted security, we can ensure that robots amplify human potential while safeguarding our shared values and physical wellbeing.

REFERENCES

- [1] C. Menon et al., "EHAZOP: A Proof of Concept Ethical Hazard Analysis of an Assistive Robot," arXiv: 2406.09239, 2024.
- [2] X. Huang et al., "Trust in LLM-controlled Robotics: a Survey of Security Threats, Defenses and Challenges," arXiv: 2601.02377, 2025.
- [3] S. Neupane et al., "Security Considerations in AI-Robotics: A Survey of Current Methods, Challenges, and Opportunities," IEEE Access, vol. 12, 2024.