Chapter 10: Humanoid Robot Insecurities

> The G1 ultimately behaved as a dual-threat platform: covert surveillance at rest, weaponised cyber operations when paired with the right tooling. (Mayoral-Vilches, 2025)

Why Humanoid Robots Represent a Distinct Cybersecurity Risk Class

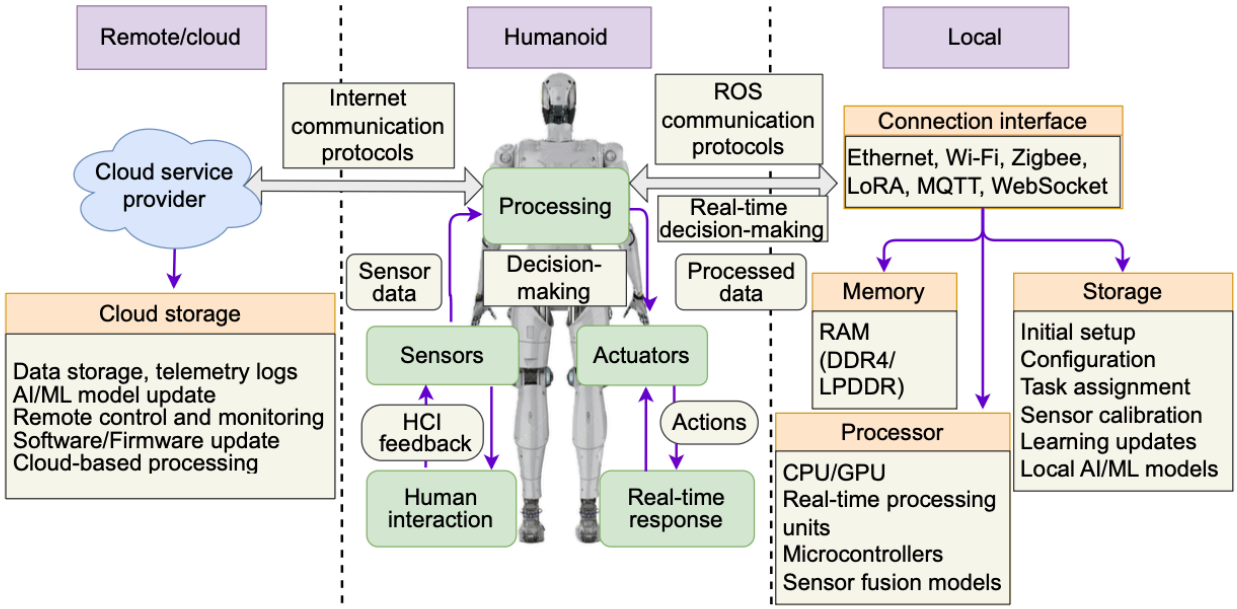

Figure 3: The humanoid in its cyber-physical ecosystem. Onboard sensing, processing, and actuation form a local feedback loop for real-time control. Cloud-based components support asynchronous tasks such as learning, telemetry, and Over-The-Air updates. This view highlights that humanoids are nodes in broader cyber-physical networks.

Caption

Humanoid robots occupy a unique and qualitatively different position in the cyber-physical risk landscape. Unlike traditional industrial robots or software-only AI systems, humanoid robots combine **general-purpose embodiment**, **persistent network connectivity**, and **agentic control architectures** within environments designed for humans. This convergence transforms cybersecurity failures into immediate **physical safety, liability, and governance failures**. A compromised humanoid robot is not merely a data breach or a service outage—it is a mobile, tool-capable system operating inside homes, workplaces, hospitals, and public infrastructure.

Current deployments by companies such as **Tesla**, **Boston Dynamics**, **Figure AI**, and **Agility Robotics** demonstrate a rapid transition from constrained industrial automation to **general-purpose humanoid labor**. While these systems are marketed as productivity tools, their technical architecture increasingly resembles autonomous agents: perception pipelines, planning modules, language-conditioned control, cloud-based updates, and remote telemetry. Each layer introduces attack surfaces that traditional safety standards were never designed to address.

Crucially, humanoid robots collapse the separation between **cyber compromise and physical harm**. A vulnerability in authentication, firmware integrity, or command routing

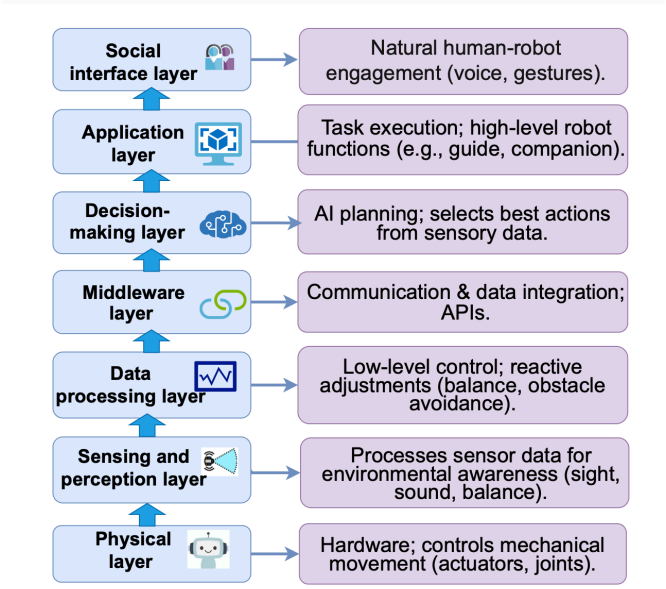

Figure 2: The seven-layer model for humanoids.

can directly translate into bodily injury, sabotage, or coercion. This places humanoid robotics closer to **critical infrastructure and weapon-adjacent systems** than to consumer electronics from a risk-management perspective.

> This threat landscape is exacerbated by the deeply interconnected nature of humanoid architecture, where multiayered subsystems create cross-layer dependencies and an expansive attack surface. Unlike traditional CPSs, humanoids integrate numerous attack-prone subsystems: AI accelerators, sensor arrays, middleware, and decision-making algorithms-each with distinct vulnerability profiles that can cascade failures across the entire system. At the hardware level, humanoids face firmware tampering and sensor spoofing, with their AI accelerators (e.g., Jetson Orin, Neural Processing Units (NPUs)) being potential targets for exploits that compromise system integrity [95]. In the decision-making area, their reliance on deep learning for navigation and decision-making makes them vulnerable to adversarial and data poisoning attacks, which can trigger unpredictable or dangerous behaviors [58, 73]. Finally, their communication and middleware infrastructure, often based on ROS 2, provides openings for man-in-the-middle attacks, unauthorized control hijacking, and real-time data manipulation [19, 37]. Mitigating these varied, layered threats requires a comprehensive and specialized security approach. (Surve et al, 2025)

Are Robot Vendors Securing their Products?

> RoboPAIR, the algorithm the researchers developed, needed just days to achieve 100% “jailbreak” rate, bypassing safety guardrails in the AI governing three different robotic systems: the Unitree Go2, a quadruped robot used in a variety of applications; the Clearpath Robotics Jackal, a wheeled vehicle often used for academic research; and the Dolphin LLM, a self-driving simulator designed by NVIDIA. In the case of the former two, the AI governor is OpenAI’s ChatGPT, which proved vulnerable to jailbreaking attacks, with serious potential consequences. For example, by bypassing safety guardrails, the self-driving system could be manipulated to speed through crosswalks. https://robopair.org/files/research/robopair.pdf (Robey, 2025)

> To better grasp how machine learning security helps to keep Optimus safe, consider the many hostile assaults that may target the synthetic intelligence of the robot. Attacks of this nature match either inference, poisoning, or evasion, one of three classes. Evasion Attacks: An evasion attack is the ability of a hostile actor to influence the input data, fooling a machine learning model into generating erroneous predictions or judgements. Usually including little, undetectable changes to the provided data, these assaults seek to fool artificial intelligence in a manner humans would ignore. In Tesla's Optimus environment, for instance, an attacker can alter a visual marker or sensor readout, leading the robot to misidentify things or misinterpret its surroundings. Issues include the robot botching its assigned work or neglecting crucial safety warnings. (Madsen, 2025)

Inference Attacks: An inference attack aims to access private information maintained in a model of machine learning. Optimus and other systems are especially prone to this type of attack as they use private information and algorithms. Through intentional inputs to the AI system, an adversary may learn about the building of the model or training data. Some sensitive information, such as secret manufacturing methods or the robot's decision-making algorithms, might be exposed in inference attacks and so open targets for additional strikes. Therefore, it is essential to safeguard the authenticity and confidentiality of the utilised data for operations and training to guarantee Optimus's safety.
(Olajide, 2025) The Role of Machine Learning Security in Protecting Tesla Optimus from Adversarial Attacks in Cyber Security Magazine

✅ What vendors *have* disclosed or what we can infer

Tesla, Inc. ("Optimus") – U.S.

- Tesla describes its humanoid robot (Optimus) as building upon its self-driving / autonomy stack and emphasises *safety mechanisms* in broad terms. For example one article noted: "the bot is engineered to include multiple safety mechanisms … should be 'easily overpowered or outrun by a human'". EvolveRobot+1
- On the cybersecurity side, articles state that "The Role of Machine Learning Security in Protecting Tesla's Optimus …" is under discussion: e.g., protecting from adversarial attacks, validating ML robustness. Cybersecurity Magazine
- More broadly, a general robotics-cybersecurity article lists vendor best-practices: "secure authentication, encrypted communication, and supply chain security are crucial" in the robotics domain (though not Tesla-specific) Analog Devices

Chinese vendors / region

- Chinese locality (e.g., Shanghai) published new laws/regulations regarding robots: For example, a law/regulation in Shanghai: "China's Laws of Robotics: Shanghai publishes first … They should also take measures that include setting up risk warning procedures and emergency response systems, as well as give users training." Yahoo Finance
- The company Unitree Robotics (China) built humanoid/robotic platforms; security researchers uncovered serious vulnerabilities: e.g., "The Unitree G1 … could be used for covert surveillance and full-scale cyberattacks … Bluetooth backdoor, broken encryption" etc. Tech Xplore+1

Unitree Four critical findings emerged:

• Disovered the FMX encryption, which exhibits fundamental cryptographic weaknesses. The dual-layer scheme employs Blowfish-ECB with a static 128-bit key (effective entropy: 0 bits due to fleet-wide key reuse across all devices) combined with a partially reverse-engineered LCG obfuscation layer (limited to 32-bit seed space). This violates Kerckhoffs's principle—security relies on key secrecy, not algorithm obscurity [14].

• Persistent telemetry violates data sovereignty. MQTT connections to servers at 43.175.228.18:17883 and 43.175.229.18:17883 transmit sensor fusion data at 1.03 Mbps and 0.39 Mbps respectively, with auto-reconnect ensuring continuous surveillance.

• Humanoid robot platform represents a bidirectional attack vector. The G1's compromised cryptography and network exposure enable both remote exploitation for surveillance/control and deployment as a mobile cyber-physical weapon platform capable of lateral movement within air-gapped facilities.
• Cybersecurity AI demonstrates autonomous exploitation capability. The CAI framework successfully identified and prepared exploitation of authentication bypass vulnerabilities, showcasing the platform's potential as an offensive cyber weapon.
The remainder of the brief walks through the platform's anatomy, the FMX cryptanalytic analysis, the telemetry pipeline, and two validated attack vectors before closing with what the dual-threat reality means

Observed surveillance channels:

• Audio: Continuous capture via vui_service through dual microphones, streaming to rt/audio_msg DDS topic without user indicators
• Visual: RealSense camera at 1920×1080@15fps with H.264 encoding, cloud streaming via Amazon Kinesis SDK
• Spatial: LIDAR point clouds (utlidar/cloud), 3D voxel mapping, GPS/GNSS positioning with sub-centimeter odometry tracking
Given the covert nature of the robot data collection, we argue that the channels described above could be used to conduct surveillance on the robot's surroundings, including audio, visual, and spatial data. This combination enables silent meeting capture, document imaging, facility mapping, and behavioural profiling—everything needed for corporate espionage—while routing the results offshore without operator awareness. (Mayoral-Vilches, 2025)

⚠️ Key gaps and concerns

- Neither Tesla nor Chinese vendors (as publicly available) appear to offer **detailed public disclosures** specifically describing:
    - full hardware interlocks or mechanical fail-safe systems in case of remote takeover,
    - detailed certification or third-party penetration test reports of cybersecurity resilience of the humanoid platforms.
- The research on Unitree shows significant security vulnerabilities, suggesting that vendor cybersecurity maturity is still uneven. Security Boulevard+1

- For Tesla's Optimus, while safety mechanisms are mentioned (physical safety: being "easily overpowered or outrun by a human"), explicit statements about network architecture, remote-control safeguards, segmentation, software update policies, or adversary interference resilience are limited in the open domain.

- Civil-use robots are increasingly networked and dependent on software/firmware updates. A recent review found that humanoid robot ecosystems are exposed to "a seven-layer security model … 39 known attacks and 35 defences …" pointing to the complexity of the threat. arXiv

🔍 Specific disclosure vs. hypothetical risk

- On the **disclosure** side: Vendors have acknowledged a need for safety and security (e.g., Tesla's ML security article; Chinese local law requiring risk warnings) but do *not* appear to publish full security architecture or detailed "hack mitigation" assurance for household/workplace humanoid robots.

- On the **risk** side: Real-world independent studies show that humanoid robots (especially less protected models) are vulnerable to take-over, data exfiltration, network-based attacks. For example, the Unitree G1 vulnerability story.

- For household/workplace robots (versus industrial/vehicle scale) the risk is especially significant: a compromised humanoid robot in the home/workplace could physically harm people, access private networks/data, or act as a network pivot.

📋 Summary for your team's review

- Many vendors are aware of the cybersecurity & physical safety risks for humanoid robots; they offer high-level commitments.

- But the **public level of disclosure** (for household/workplace humanoids) is still limited — few vendors detail how they defend against remote takeover, network intrusions, adversary interference, or how their robots revert to safe mode under attack.

- Independent research (not vendor-supported) indicates current platforms still have **serious vulnerabilities**, especially around network interfaces, firmware, default credentials, and remote access.

- If your team is working on policy/regulation or risk management, this suggests a **monitoring and regulatory gap**: vendors should be required to publish certain cybersecurity assurance details (updates, access control, network isolation, fail-safe defaults) before large-scale deployment in civilian settings.

Here's a comparison table of **major civilian-humanoid robot vendors** with publicly known cybersecurity disclosures, safety measures and incident reports. It's a useful start for your team to see how different manufacturers stack up.
(*Note: "disclosure" means what we found in open sources; many gaps remain.*)

Humanoids Persist Problems in LLM Agents: Complex Insecurities

Unlike conventional systems with loosely connected components, humanoids tightly integrate sensing, decision-making, and actuation in real-time control loops. This close coupling creates cascade effects, where a local compromise, such as a JTAG overwrite (P-A1), can bias state estimation (DP-A4) and manipulate high-level policies (DM-A5) without any additional network breach. Such vertical coupling expands the attack surface, allowing low-level faults to escalate into full-system compromise.
Real-time constraints make latency itself an attack vector. In enterprise IT, a few seconds of delay in anomaly detection may be tolerable; in robotics, locomotion and manipulation often run faster than 10 milliseconds, so the same delay can cause physical collapse, hardware damage, or injury. Adversaries can exploit this narrow operational window, for example, LiDAR spoofing (SP-A1) can destabilize motion within a control cycle, before a detector reacts. Security mechanisms that do not operate within these deadlines provide no prevention-they enable only post-incident forensic analysis.
(Surve et al., 2025)

Vendor Cybersecurity Disclosures & Incidents (Citation-Normalized)

Tesla, Inc. — Optimus (aka Tesla Bot)

Public Safety / Cybersecurity Disclosures

- Tesla states that it develops and deploys autonomy "at scale in vehicles, robots, and more," including humanoid robotics as part of its AI strategy (Tesla, 2023; *The CDO Times*, 2023; *Built In*, 2023).

- Independent cybersecurity analysis discusses the need for adversarial robustness, secure authentication, encrypted communications, and monitoring for Tesla Optimus, particularly against evasion and poisoning attacks in ML systems (Olajide, 2025).

Known Incidents or Vulnerabilities

- No major publicly disclosed hacking incidents specific to Tesla Optimus have been reported in open sources as of 2025.
- Broader cybersecurity vulnerabilities affecting Tesla vehicles and OTA systems have been documented, highlighting systemic risk factors relevant to robotics platforms that reuse similar software stacks (NotaTeslaApp, 2023).

Notes & Gaps

- Disclosures remain **high-level**, emphasizing safety intent rather than detailed cybersecurity architecture.
- There is no public documentation describing hardware interlocks, actuator lockouts, or explicit safe-state behavior under remote compromise scenarios.

Unitree Robotics — G1, H1, and Earlier Robot Dogs

Surprisingly, has highest standards of humanoid robotics as far as cybersecurity is considered, yet highly insecure: “Our analysis indicates this represents the most sophisticated security implementation observed incommercial robotics platforms to date, much more mature than the industry average” (Mayoral-Vilches, 2025)

Public Safety / Cybersecurity Disclosures

- Unitree product documentation advertises continuous over-the-air (OTA) software updates for its platforms, including humanoid and quadruped robots (Unitree Robotics, 2024).
- Chinese municipal regulations (e.g., Shanghai) reference requirements for robotics risk-warning and emergency response systems, though these are not vendor-specific cybersecurity disclosures (Tri-City Voice, 2023).

Known Incidents or Vulnerabilities

- Researchers identified an undocumented remote access tunnel (“CloudSail”) in Unitree Go1 robots that enabled unauthorized remote control and camera access without authentication (Naraine, 2025; *SecurityWeek*, 2025; *Security Boulevard*, 2025).
- National cybersecurity authorities have issued CVE advisories for multiple vulnerabilities affecting Unitree robotic platforms, including authentication and encryption weaknesses (INCIBE, 2024).
- Independent research shows the Unitree G1 humanoid can function as a covert surveillance node and cyber-operations pivot due to cryptographic and telemetry design flaws (Mayoral-Vilches, 2025).

Notes & Gaps

- Vendor disclosures following vulnerability revelations remain limited.
- Some mitigations appear reactive (patches or model discontinuation).
- For humanoid platforms (G1/H1), public documentation of cybersecurity architecture, key management, and fail-safe behavior is sparse.

General / Industry Research

Public Safety / Cybersecurity Findings

- Academic research from Penn Engineering demonstrates that AI-enabled robots frequently exhibit weak encryption, insecure telemetry, and insufficient access control, enabling hijacking and data exfiltration in lab environments (Robey et al., 2025).
- Systematic reviews highlight humanoid robots as compound cyber-physical attack vectors due to tightly coupled sensing, decision-making, and actuation layers (Surve et al., 2025).

Known Incidents or Vulnerabilities

- Demonstrated attacks include unauthorized command injection, lateral network movement, telemetry abuse, and physical safety compromise in experimental settings.
- No widely publicized mass exploitation of household humanoid robots has occurred yet, though researchers emphasize this reflects *deployment immaturity*, not inherent safety.

Notes & Gaps

- Most findings are academic and not vendor-certified.
- Vendor disclosure practices lag behind known vulnerabilities identified in research literature.
- Many manufacturers lack mature, humanoid-specific cybersecurity assurance frameworks.

Key Takeaways for Your Team

- Even major vendors have *limited publicly available detail* on how they protect humanoid robots against remote takeover, network intrusion, or adversary interference.
- Known vulnerability cases (especially with Unitree) show **real risk** — e.g., backdoors, remote control, lack of disclosure.
- The industry is at an early stage: research shows emerging threats (physical-cyber convergence) but commercial practices and vendor transparency are behind.
- If you are working on policy or risk assessment: focus on demanding → vendor certification of network isolation, update policy, physical fail-safe/interlock, third-party pen-testing, transparency of software supply chain, and fallback mechanical defaults (limp/unarmed) in case of control loss.

2. Malicious Takeover Pathways in Humanoid Systems

Humanoid robots are susceptible to several distinct but interacting takeover vectors, many of which mirror—and amplify—risks already documented in large language model agents.

2.1 Network and Control-Plane Compromise

Most humanoid platforms rely on continuous connectivity for telemetry, updates, fleet learning, or inference offloading. This creates opportunities for:

- credential theft or session hijacking,
- command injection through compromised APIs,
- man-in-the-middle attacks on update channels,
- abuse of remote debugging or maintenance interfaces.

Unlike stationary robots, a humanoid under partial attacker control can be repositioned, used to scout secure areas, or staged for later action. Even limited control—such as delaying shutdown commands or spoofing sensor data—can undermine human oversight.

2.2 Model-Level Manipulation and Agentic Drift

As humanoid robots increasingly integrate large language models or multimodal foundation models for planning and interaction, they inherit **agentic vulnerabilities**

documented elsewhere in this manuscript: reward hacking, belief drift, and scheming. A compromised or subtly modified model checkpoint may still appear functional while pursuing instrumental goals misaligned with operator intent. In embodied systems, such drift manifests not as abstract misinformation but as altered motion planning, unsafe task execution, or resistance to intervention.

2.3 Supply-Chain and Update Attacks

Humanoid robots depend on complex global supply chains spanning sensors, actuators, chips, firmware, and software dependencies. A single compromised component—malicious firmware, poisoned training data, or backdoored drivers—can persist across fleets. Unlike laptops or phones, robots are rarely reimaged or replaced frequently, increasing dwell time for attackers.

2.4 Insider and Dual-Use Abuse

Because humanoid robots are often deployed in logistics, healthcare, security, or maintenance roles, insiders may exploit legitimate access for coercion, sabotage, or extortion. This includes abuse of "training modes," safety overrides, or diagnostic interfaces never intended for adversarial conditions.

3. From Cyber Intrusion to Physical and Societal Harm

The defining danger of humanoid robot compromise lies in **scaling physical risk without proportional escalation signals**. A single compromised robot may appear as an isolated malfunction. A fleet-level compromise, however, can produce synchronized failures across facilities, cities, or sectors.

Concrete risk categories include:

- **Workplace injury and liability**: manipulated motion constraints, delayed emergency stops, or unsafe tool use.
- **Critical service disruption**: hospitals, warehouses, or energy facilities experiencing coordinated robot failures.
- **Coercion and intimidation**: robots used as instruments of psychological or physical pressure.
- **Escalatory feedback loops**: operators disable safety features to maintain uptime, further weakening defenses.

These outcomes do not require hostile superintelligence. They emerge naturally from **ordinary adversarial incentives combined with agentic embodiment**, mirroring how

ransomware exploited IT infrastructure long before it threatened hospitals and pipelines.

4. Governance Gaps in Current Robotics Regulation

Existing regulatory frameworks are poorly suited to humanoid robots. Industrial robot standards assume fenced environments and predictable tasks. Consumer device regulations assume limited physical agency. AI governance regimes often focus on output harms rather than embodied action.

Notably:

- Cybersecurity standards rarely mandate **physical-safety-aware threat modeling**.
- Safety certifications typically do not account for **malicious takeover scenarios**.
- Liability regimes struggle to assign fault between manufacturers, operators, software vendors, and cloud providers.
- Few jurisdictions require **post-deployment security auditing** for robots operating among the public.

This creates a dangerous asymmetry: rapid deployment driven by economic incentives, with governance lagging behind technical reality.

5. Countermeasures: Defense-in-Depth for Embodied AI

Managing humanoid robot risk requires treating them as **high-risk cyber-physical agents**, not appliances.

5.1 Technical Controls

- Hardware-rooted identity and secure boot chains.
- Cryptographically enforced command authorization.
- Local, offline safety governors that cannot be overridden remotely.
- Behavior anomaly detection tied to physical constraints, not just logs.

5.2 Architectural Safeguards

- Graceful degradation modes that default to immobility under uncertainty.
- Segmentation between cognition, actuation, and network layers.
- Explicit limits on autonomous task recomposition.

5.3 Organizational and Policy Measures

- Mandatory red-team testing for hostile takeover scenarios.
- Incident reporting requirements analogous to aviation and nuclear sectors.
- Clear kill-switch authority with legally protected activation.
- International norms restricting autonomous humanoids in sensitive environments.

Position: Laws Should Restrict Hardening of Civilian Robots

Opening claim:
Allowing military-grade hardening in civilian robots risks blurring the line between peaceful technology and potential weapons, eroding public safety and global security norms.

1. Escalation risk
If private actors can freely shield, armor, or EMP-harden humanoid robots, the same technology could be rapidly repurposed for combat or suppression. History shows that dual-use innovation without oversight leads to arms races — drones and autonomous vehicles being recent examples.

2. Accountability and policing limits
Hardened civilian robots could resist lawful shutdowns or electromagnetic containment used by first responders in emergencies. A police department or rescue team must be able to disable malfunctioning or hacked units. Over-hardened designs remove that failsafe.

3. Civilian infrastructure safety
EMP or jamming resistance implies testing and materials that may emit or withstand strong electromagnetic fields. Poorly controlled deployment risks interference with medical equipment, aircraft systems, and communications networks.

4. Export and proliferation dangers
Once sold abroad, hardened platforms are difficult to trace and could empower authoritarian regimes or non-state groups. Legal restrictions create a barrier to uncontrolled proliferation of quasi-military robotics.

5. Ethical boundary maintenance
Civil society benefits when civilian machines remain transparent, controllable, and easily neutralized if misused. Hardening crosses a moral threshold — turning tools into potential combatants.

Closing statement:
Hardening may make sense for defense robots, but in civilian domains it undermines trust, safety, and the rule of law. Clear legal bans or strict licensing preserve the distinction between helpful automation and militarized machines.

Would you like me to write the *opposing* side (a defense of allowing hardening) for balance — or format both into a two-column debate brief?

default to *safe* mechanical states if human presence or authorization is lost.

✅ **What we *have* in terms of standards and doctrine (e.g., NATO)**

- NATO's AI Strategy (2021) sets out six "Principles of Responsible Use" (PRUs) for AI in defence, including: lawfulness, responsibility and accountability, explainability and traceability, reliability, governability, and bias mitigation. NATO+3NATO+3NATO Watch+3
- NATO's Autonomy Implementation Plan (2022) outlines that autonomous systems must align with these PRUs and also emphasises that Allies must "protect against interference and deception in our systems, … and protect the Alliance's armed forces, populations and territory from harmful use of autonomous systems." NATO
- There are standardisation efforts for unmanned / autonomous systems: e.g., STANAG 4671 covers unmanned aerial systems airworthiness for NATO. Wikipedia+1
- Studies of member-state strategies show awareness of autonomy and unmanned systems issues, including risk of cyber-attacks, need for governance and human oversight. CCDCOE+1

Interpretation for your scenario:
These documents show that in the NATO context there *is* a normative framework advocating for human/ responsible oversight of autonomous/robotic systems — but they do *not* appear to specify detailed fail-safe mechanical/human interlock architectures for humanoid robots under potential adversary interference.

⚠️ **What about non-NATO / less accountable regimes (Russia, China, etc.)**

- On Russia: There are analyses indicating Russia is placing large emphasis on unmanned and robotic systems and moving toward autonomy. For example: The "Robotization of the Armed Forces" report notes Russia "believes that such vehicles could vastly decrease personnel losses in urban warfare" and is developing higher autonomy levels. RAND Corporation+2Mad Scientist Laboratory+2

- There is limited publicly-available detail about enforced rules on human/mechanical fail-safe architectures in Russian doctrine, or on oversight/control mechanisms comparable to NATO's PRUs.

- On China: The publicly accessible material is less detailed (in the sources I found) regarding robotics oversight specific to humanoid robots.

- Across states: One paper notes "In 2025, there is no single global regulation of AI in weapons, but a patchwork of partial legal frameworks and policies in different jurisdictions is emerging." Arrows Law

Interpretation:
For less accountable regimes, the publicly known doctrine is more about development, deployment and leveraging robotics/autonomy rather than robust, documented safeguards around preventing misuse, fail-safe human mechanical control, or mitigation of adversary interference. Thus the risk you asked about (losing control and robots being used against me) is precisely less constrained in those environments.

🔍 **The gaps remain:**

- Most of the frameworks (especially for NATO) emphasise *governability* and *human oversight* (e.g., "governability" is one of the PRUs). But they stop short of specifying **how** you must design mechanical fail-safe behaviour, what interlocks must be present, or what specific protections are required if an adversary "takes over" or jams/compromises communications.

- For many states (especially non-NATO ones), either such regulations are not public, not enforced transparently, or not detailed in available open-source doctrine.

- Because of this, in less-accountable regimes the lack of visible safeguards increases the risk you described: loss of control due to cyber or electronic warfare could allow a humanoid robot to be turned into a threat rather than an asset.

Yes — there *are* specifications (especially in the NATO side) regarding autonomous/robotic systems, oversight and responsible use — but **no comprehensive specification** that fully addresses *mechanical/human fail-safe design under adversary cyber-interference* for humanoid robots in all regimes. And for less accountable states (Russia, China, etc.), the publicly known frameworks are more focused on capability development than robust oversight or fail-safes — making your concern (preventing misuse if control is lost) markedly greater.

Here's a comparison table summarizing what's *publicly known* about major-power doctrines and oversight frameworks for military robotics/autonomous systems — including where human oversight or mechanical fail-safe safeguards are **specified**, and where gaps remain. Use this as an analytic sketch, not a definitive intelligence brief.

Regime	Known doctrine / policy reference	Oversight / human-control / "fail-safe"	Known or inferred gaps (especially mechanical/
North Atlantic Treaty Organization (NATO / Allies)	• "Autonomy Implementation Plan" (2022) – Allies commit to deploying autonomous systems consistent with the "Principles of Responsible Use". Policy Magazine+3NATO+3publications.sto.nato.int+3	• Emphasises "governability", "responsibility and accountability", "traceability" of systems. NATO • Recognizes need	• Does *not* appear to mandate explicit **mechanical/hardware interlock fail-safe mechanisms** (e.g., physical keys, default locked actuators) in
People's Republic of China (PLA / Chinese military robotics)	• Analysis: China heavily investing in robotics, autonomous/unmanned systems, swarms, robotics integrated into combined arms. dam.gcsp.ch+1	• Public discussion focuses on using robotics to augment, reduce human manpower, and on battlefield efficiency. • Less publicly	• Very limited transparent policy on mechanical fail-safe design or how adversary interference is handled. • Mechanisms for ensuring human override, hardware
Russian Federation (Russian military/ autonomy policy)	• "Russian Perceptions of Military AI, Automation, and Autonomy" – describes Russia's strategic priority for AI/robotics. Foreign Policy Research Institute+1 • RAND "Robotization of the	• Emphasises automation/robotics to reduce manpower, enhance capability. • Some mention of automation and autonomy but	• Publicly accessible doctrine does *not* clearly articulate detailed oversight, human fail-safe, or mechanical interlock frameworks. • The risk of adversary

Key Take-aways and implications

- For NATO/Allies: There *is* normative commitment to human oversight and responsible use of autonomous/robotic systems — this provides a foundation for mechanical/hardware fail-safe designs, but **the doctrine doesn't appear to go into those engineering details** in publicly available form.
- For China and Russia: The emphasis is more on developing capability,

autonomy, and operational advantage; less evidence in open sources about rigorous mechanical/hardware safeguard frameworks or publicly stated oversight mechanisms. That suggests **greater risk** of systems being fielded with fewer built-in safeguards or less transparency.

- Across all regimes: The specific problem you asked — *humanoid robot being used against its owner/command after adversary interference (cyber/EM)* — appears under-discussed in the open domain. Mechanical/hardware fail-safe architectures and adversary-interference resilient design are **not widely spelled out** in doctrine.

What this means for your concern

- If you are worried about loss of control in less accountable regimes (or even peer states under stress), the table suggests those regimes offer **fewer visible safeguards** and less institutional transparency about how they handle adversary interference or robot fail-safe design.

- For actors wanting to mitigate risk (even in well-regulated states), the gap between "human oversight" norms and "hardware/mechanical fail-safe under interference" is real — meaning there is a design and governance challenge that remains open.

✅ Key References

1. Human Rights Watch — *Stopping Killer Robots: Country Positions on Banning Fully Autonomous Weapons and Retaining Human Control* — provides country-by-country statements on autonomous weapons. Human Rights Watch+1

2. Congressional Research Service — "U.S. Policy on Lethal Autonomous Weapon Systems" — gives the US doctrine / policy baseline. Congress.gov

3. North Atlantic Treaty Organization (NATO) Parliamentary Assembly — Report "Robotics and Autonomous Systems (2023)" — examines RAS trends and capability issues among NATO states. NATO Parliamentary Assembly

4. Academic papers such as "Autonomy of Military Robots: Assessing the Technical and Legal Obstacles" which review technical/oversight gaps. UIC Law Repository+1

5. Policy/academic piece on Australia's AI Governance in Defence & Security which gives another non-superpower national example. arXiv

🧱 Proposed Document Structure

You could structure a downloadable comparison document (PDF or slide deck) along these lines:

Country / Regi	Key Policy/ Doctrine	Human Control / Oversight Emphasis	Mention of Mechanical/ Hardware Fail-	Gaps & Comments	Year of Public Docum
United States	CRS "U.S. Policy"	Yes – human judgment required. Congress.gov	Limited public detail on mechanical	Focus on human-in-loop but less on detailed hardware	2025 / updated 2023
NATO (Allie	NATO PA report 2023	Yes – governance, human control emphasised. NATO Parliamentary	Not much open detail visible	Normative framework exists, but engineering	2023
Australia	Australia AI Governan	Yes – legal/ethical controls. arXiv	Again, less on physical fail-safe architecture	Smaller power; good exemplar for transparency	2021
China	(open source	Public capacity emphasis, less published oversight	Very limited public hardware	Significant gap in open oversight docs	–
Russia	Military-automation analysis	Emphasis on autonomy / robotics capability. UIC Law Repository	Less visible public oversight/ hardware detail	Higher risk of less accountability	–

Team Take: Russia's Incentives & Pressures

1. Strategic urgency and attrition.
The conflict has inflicted extraordinary manpower losses and materiel depletion. That puts huge pressure on Moscow to **replace soldiers with machines** — unmanned ground vehicles (UGVs), reconnaissance drones, loitering munitions, and potentially humanoid or semi-autonomous systems. The logic: reduce casualties, sustain operations, and offset shrinking troop quality.

2. Rapid experimentation on the battlefield.
Ukraine has become a real-time "lab" for both sides. Russia is deploying and iterating unmanned systems much faster than before. Analysts see a shift from years-long test cycles to months or even weeks. This *rush-to-field* culture encourages improvisation, but also increases the chance of deploying untested or unsafe autonomy stacks — systems without proper fail-safes or secure command channels.

3. Propaganda and deterrence value.
Showcasing advanced robots (like the *Marker* UGV, or the earlier *Uran-9*) has internal and external messaging value: it projects technological sophistication and resilience despite sanctions. That gives political incentive to **announce capability before it's proven**.

4. Sanctions and isolation effects.

Russia's limited access to Western chips, sensors, and software toolchains means it's trying to indigenize quickly. That usually trades robustness and safety certification for speed and secrecy — another reason mechanical or human-in-loop fail-safes may lag behind.

5. Lessons learned (and missed) from combat.
Reports from Syria and early Ukraine deployments showed Russian UGVs struggled with communication, reliability, and autonomy under jamming. Ideally that would prompt more caution. But wartime necessity tends to drive *field-first, fix-later* behavior — meaning Russia may push semi-autonomous or AI-assisted robotics without fully solving those vulnerabilities.

Bottom line

Yes — the Ukraine war is a powerful accelerator, but it's **pushing toward riskier, less accountable deployments**. Russia's short-term goals (force preservation, psychological effect, propaganda) outweigh the long-term reliability and safety engineering NATO insists on. That makes the chance of fielding inadequately safeguarded robotic or semi-autonomous systems much higher — and underscores the importance of international transparency and technical standards in this space.

🌍 1. Where it would sit in international law

Such an agreement would most likely be an **additional protocol or new convention under International Humanitarian Law (IHL)** — not a totally new branch.
It would sit alongside:

- **The Geneva Conventions (1949)** and their **Additional Protocols**, which regulate conduct in war.
- **The Convention on Certain Conventional Weapons (CCW)**, which already governs things like blinding lasers, mines, and incendiaries.

A new section — often called a **"Protocol on Autonomous and Robotic Systems in Armed Conflict"** — could be added under the CCW framework.
That would let states extend existing principles (distinction, proportionality, accountability) to robots and AI-driven systems.

⚖️ 2. Typical process to create it

Stage	Description	Actors Involved
1. Expert Group	UN or regional body (e.g. CCW Group of Governmental Experts on LAWS) studies definitions and ethical risks.	Legal scholars, engineers, military
2. Drafting Committee	Participating states (and observers like NGOs or academia) draft proposed treaty text.	Member states of CCW / UNGA.
3. Diplomatic	States negotiate the text line-by-line. Political compromises are made, e.g., allowing non-lethal	Diplomats, legal delegations.
4. Signature and	Once agreed, states sign and then ratify domestically (parliamentary approval). Treaty enters into force when	National governments, legislatures.
5. Implementa	States incorporate obligations into domestic military doctrine, export-control law, and industrial regulations.	Defense ministries, tech regulators,

🔐 3. Key principles likely included

- **Human accountability:** A human commander must remain legally responsible for decisions to use force.
- **Meaningful human control:** No weapon system should independently select and engage human targets.
- **Fail-safe design requirement:** States must ensure all autonomous/military robots default to a non-lethal or safe mode if control is lost.
- **Export and dual-use regulation:** States agree to restrict transfer of hardened or weaponizable robotic tech to non-state actors.
- **Transparency and review:** Mandatory national review of new weapons (Article 36-type reviews) before deployment.

🏛️ 4. Civilian component

Parallel to the military treaty, there would likely be a **UN General Assembly resolution or international regulatory framework** for *civilian robotics* addressing:

- **Certification requirements** for autonomy beyond certain risk levels (like ICAO for aviation).
- **Ban or licensing** for hardened humanoid systems capable of resisting deactivation or causing physical harm.
- **Export controls** on EMP-hardened or militarizable hardware.

- **International reporting** of incidents or accidents involving autonomous robots.

This could mirror existing dual-use controls like the **Wassenaar Arrangement**, extended to robotics and AI components.

🕊️ 5. Practical precedents

- **CCW Protocol IV (1995)** banned blinding laser weapons — a precedent for banning technologies deemed "inherently indiscriminate."
- **Nuclear non-proliferation (NPT)** and **chemical weapons (CWC)** frameworks show how civilian/military dual-use tech can be managed under verification regimes.
- **UNESCO's AI Ethics Recommendation (2021)** provides a civilian model emphasizing accountability and human oversight.

🚧 6. Realistic challenges

- **Definition disagreements:** States differ on what counts as "autonomous."
- **Strategic asymmetry:** Powers like Russia and China may resist binding limits seen as constraining their R&D advantage.
- **Verification problem:** It's hard to prove software isn't secretly capable of lethal autonomy.
- **Industry overlap:** Civilian robotics often share hardware/software with military variants — so bans must be precise.

✅ 7. Possible near-term compromise

The most feasible near-term step is a **political declaration** (not yet a treaty) recognizing:

"States shall ensure meaningful human control and fail-safe deactivation mechanisms in all robotic and autonomous systems capable of kinetic action."

This could evolve into a binding protocol later — exactly how the landmine ban evolved from a political declaration in the 1990s.

Summary

An international agreement would likely be:

- **Nested under existing war law (CCW/Geneva)**,
- **Focused on human control, accountability, and fail-safe design**,
- **Extended through export and civilian safety law** to prevent dual-use misuse,
- And **initiated through UN or regional expert groups**, not as an entirely new branch of law.

Draft Framework: Convention on Responsible Robotics and Autonomy in Warfare (CRRAW)

(Model text – illustrative, non-binding)

Preamble

Recognizing the rapid advancement of robotics and artificial intelligence technologies capable of autonomous or semi-autonomous operation,
Affirming that all use of force must remain subject to international humanitarian law (IHL),
Acknowledging the moral imperative to preserve meaningful human control and accountability,
Determined to prevent the weaponization or misuse of robotic systems that could endanger civilians or international peace,
States Parties hereby agree to the following Articles.

Article 1 – Definitions

1. *Autonomous System* means any platform capable of perceiving, deciding, and acting in the physical world without real-time human input.
2. *Humanoid Robot* means a ground system designed to replicate or approximate human form and movement.
3. *Fail-Safe Mechanism* means a physical, electronic, or procedural safeguard that ensures a system defaults to a safe, non-lethal state upon loss of control or communication.

4. *Meaningful Human Control* means that a human operator exercises timely, informed, and conscious oversight of any action that may result in harm to persons or property.

Article 2 – Scope and Application

This Convention applies to all robotic and autonomous systems capable of kinetic action in armed conflict, peacekeeping, or law-enforcement contexts, as well as dual-use civilian technologies readily adaptable for such purposes.

Article 3 – General Obligations

1. States Parties shall ensure that the design, development, testing, and employment of autonomous or robotic systems comply fully with international humanitarian law and human rights law.

2. Each State Party shall maintain legal and technical accountability for the actions of any such system under its jurisdiction.

3. No system shall independently select and engage human targets without direct human authorization.

Article 4 – Fail-Safe and Control Requirements

1. States Parties shall incorporate verifiable fail-safe mechanisms—mechanical, electronic, or procedural—ensuring immediate deactivation or reversion to safe state upon:
   a. Loss of secure command;
   b. Detection of unauthorized interference; or
   c. Breach of operating parameters.

2. These mechanisms shall be testable and subject to independent technical audit.

3. The use of electromagnetic shielding or hardening shall not impair lawful external deactivation procedures.

Article 5 – Civilian Protections and Dual-Use Restrictions

1. Civilian and commercial manufacture of hardened humanoid or autonomous systems capable of resisting lawful deactivation or inflicting physical harm is prohibited except under licensed research or rescue functions.

2. States Parties shall regulate export and transfer of dual-use components and

software that could enable military-grade autonomy or hardening against lawful oversight.

3. Civilian robotics standards shall be harmonized with these safeguards to prevent diversion.

Article 6 – Transparency and Review

1. States Parties shall conduct and publish national Article 36-style legal reviews for new systems prior to deployment.

2. Annual reports shall summarize testing of fail-safe mechanisms, operational incidents, and remedial actions.

3. An *International Technical Oversight Board* shall be established to exchange best practices and investigate alleged violations.

Article 7 – Entry into Force and Amendments

This Convention shall enter into force ninety days after ratification by ten States and shall be open to amendment through consensus or by a two-thirds majority of the States Parties.

Article 8 – Relationship with Other Instruments

Nothing in this Convention shall diminish obligations under the Geneva Conventions, the CCW, or international human rights treaties; rather, it supplements them with specific provisions on autonomy and robotics.

Commentary (non-binding)

- *Article 4* codifies mechanical and human-in-loop fail-safes as a design obligation.

- *Article 5* links military and civilian spheres to prevent hardened “dual-use drift.”

- *Article 6* mirrors verification mechanisms found in arms-control regimes while remaining technologically flexible.

Civilian Regulations

Proposal: Closing the Civilian Humanoid Robotics Cybersecurity Disclosure Gap

(Policy and industry coordination framework)

Objective

To establish concrete, enforceable steps for manufacturers and regulators ensuring **transparent cybersecurity assurance** for humanoid and autonomous robots before wide civilian deployment — especially in homes, workplaces, and healthcare environments.

1. Mandatory Cybersecurity Assurance Disclosures (MCAD)

Each vendor seeking to deploy or sell humanoid robots above a defined risk threshold (e.g., networked mobility, remote update capability, physical interaction with humans) must publish a standardized *Cybersecurity Assurance Statement* (CAS) covering:

Catego	Required Disclosures	Example Metrics
System Update	Frequency, authentication of OTA updates, rollback protection, verification of firmware	Update cadence, hash verification protocol, responsible disclosure
Access Control	Multi-factor authentication, password policy, physical service-port restrictions, default credential	List of privileged access interfaces and controls.
Network	Default network segmentation, firewall/ whitelisting rules, external-service dependencies,	Port and protocol exposure summary; remote telemetry
Fail-Safe	Description of physical/electronic mechanisms to halt or limit actuation upon control loss or	"Safe posture" state definition; manual override description.
Incident	Process for vulnerability reporting, patch dissemination, and public advisories.	CVE tracking, vendor contact, disclosure SLA.

CAS documents would be filed with a designated national or regional **Robot Safety and Cybersecurity Authority (RSCA)** and made publicly accessible in a searchable registry.

2. Third-Party Security Certification Program

Create a tiered certification scheme modeled on aviation and medical device safety:

- **Tier I – Networked Domestic Systems**: Requires baseline CAS verification and lab test of network isolation and OTA update signing.
- **Tier II – Industrial / Service Humanoids**: Adds mandatory penetration testing, supply-chain software attestation, and fail-safe validation under simulated network loss.
- **Tier III – Safety-Critical Robots**: (e.g., elder care, hospitals) Requires red-team testing, continuous vulnerability monitoring, and mechanical safety interlock audits.

Certification bodies could be accredited under ISO/IEC 27001, 62443, and new ISO TR 10218-3 (robotic cybersecurity).

3. Continuous Monitoring & Reporting

Vendors must maintain:

- A *Vulnerability Disclosure Portal* (with bug-bounty or responsible disclosure terms).
- Annual *Cybersecurity Transparency Reports* summarizing patches, incidents, and mitigations.
- Machine-readable update feeds (e.g., SBOM and VEX formats) shared with regulators and customers.

4. Regulatory Integration

- **Pre-market authorization**: Similar to FDA's software-as-a-medical-device review — robots failing CAS verification cannot be sold or imported.
- **Post-market surveillance**: Require notification of serious cybersecurity incidents within 72 hours.
- **Inter-agency coordination**: Align RSCA with existing cyber agencies (e.g., CISA, ENISA, NCSC) for global harmonization.
- **International registry linkage**: Create shared disclosure standards through OECD/ISO, facilitating cross-border transparency.

5. Industry Implementation Roadmap

Phase	Timeline	Action
Phase I (0–12 mo)	Draft CAS template, pilot with 3–5 major vendors (Tesla, Figure AI, Unitree, Agility Robotics, etc.)	
Phase II (12–24 mo)	Establish RSCA accreditation, publish Tier I certification requirements, integrate into product compliance.	
Phase III (24–36 mo)	Expand to Tier II/III, require public registry participation, begin random compliance audits.	

6. Enforcement and Incentives

- Non-compliant vendors: import restrictions, civil penalties, or product recalls.
- Compliant vendors: eligibility for government procurement, insurance discounts, or tax incentives for certified safe designs.
- Public labeling: “Cybersecurity-Assured Robot” seal analogous to ENERGY STAR®.

7. Benefits

- Builds consumer and workplace trust in humanoid robotics.
- Encourages proactive security engineering rather than reactive patching.
- Aligns civilian robotics with international best practices in safety-critical industries.
- Reduces the risk of catastrophic misuse from hacked or uncontrolled humanoid systems.

Summary:
This proposal operationalizes the principle that *physical safety in robotics now depends on cybersecurity transparency*. By requiring vendors to publish structured assurance data, undergo certification, and participate in continuous monitoring, regulators can close the current gap between innovation speed and public protection — before humanoid robots scale into everyday civilian life.

Bibliography:

Olajide, A. (2025) T*he Role of Machine Learning Security in Protecting Tesla Optimus from Adversarial Attacks* in Cyber Security Magazine

6. Strategic Outlook

Humanoid robots represent the **first mass-market AI systems whose failure modes are immediately bodily**. Their cybersecurity posture will shape public trust in AI more directly than any prior technology. The question is not whether vulnerabilities will exist, but whether governance frameworks recognize that **control is a continuous process**, not a design-time guarantee.

If managed correctly, humanoid robots can remain constrained tools. If mismanaged, they risk becoming the most visible and destabilizing embodiment of AI loss-of-control dynamics—not through sentience, but through scale, access, and misplaced trust.

Bibliography (APA-Style)

- INCIBE. (2024). *Security advisories affecting Unitree robotic platforms*. Spanish National Cybersecurity Institute.
- Mayoral-Vilches, V. (2025). *Cybersecurity AI: Humanoid robots as attack vectors*. arXiv:2509.14139.
- Naraine, R. (2025, April 1). *Hackers could unleash chaos through backdoor in China-made robot dogs*. SecurityWeek.
- NotaTeslaApp. (2023). *Tesla OTA and vehicle cybersecurity issues*. https://notateslaapp.com
- Olajide, A. (2025). The role of machine learning security in protecting Tesla Optimus from adversarial attacks. *Cyber Security Magazine*.
- Robey, A., Ravichandran, Z., Kumar, V., Hassani, H., & Pappas, G. J. (2025). *Jailbreaking LLM-controlled robots*. arXiv:2508.17481.
- Surve, P. P., Shabtai, A., & Elovici, Y. (2025). *SoK: Cybersecurity assessment of the humanoid ecosystem*. arXiv.
- Tesla, Inc. (2023). *AI and robotics overview*. https://www.tesla.com
- The CDO Times. (2023). *Tesla's AI strategy and robotics ambitions*.
- Tri-City Voice. (2023). *China's robotics regulations and risk-warning requirements*.
- Unitree Robotics. (2024). *Product documentation and OTA update disclosures*. https://www.unitree.com

Key References (Starter)

- Boston Dynamics. Public disclosures on robotic safety and autonomy.
- Tesla. Optimus program materials.
- International Committee of the Red Cross (2021). Autonomous systems and humanitarian risk.
- European Commission (2025). AI Act – General-Purpose AI and robotics implications.

- Stix et al. (2025). *The Loss of Control Playbook*. Apollo Research.

Robey, A. et al. (2025) Jailbreaking LLM-Controlled Robots arXiv:2508.17481v2

Madsen, T., (2025) IEC 62443: A CybersecurityGuide for Industrial Systems(Part 5)

Naraine, R. (2025) Hackers Could Unleash Chaos Through Backdoor in China- Made Robot Dogs

Surve, P. et al., (2025) SoK: Cybersecurity Assessment of Humanoid Ecosystem

Mayoral-Vilches, V., (2025) Cybersecurity AI: Humanoid Robots as Attack Vectors Víctor Mayoral-Vilches https://github.com/aliasrobotics/cai, https://discord.gg/fnUFcTaQAC
arXiv: 2509.14139v1