Chapter 10: Humanoid Robot Complex Insecurities

> "We should be very careful about AI. If I were to guess what our biggest existential threat is, it's probably that."
> — **Elon Musk**, National Governors Association (2017)

It's a bit cliche at this point, the whole killer robots take out humanity Hollywood depictions that become ingrained into our collective conscious, even though the evil persona robot character is very tempting, we have already seen how such an advanced intelligence would not be necessary for catastrophic Hollywood movie plots. Why many hyperbolize comments by Elon Musk about maintaining control of a Robot Army, he had a wise insight several years before, which puts such statements in perspective, one need not agree with the need to re-educate people to a certain ideology to see how information cycles are becoming extreme in every sense. So, sparing any "I, Robot" plot rehashes what is it about humanoid robots that could be threatening, in real computer security terms?

Why Humanoid Robots (HR) Represent a Distinct Cybersecurity Risk Class

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, kinda like having a industrial self-driving fork lift in your living room, with incumbent industrial hazards. 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 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.

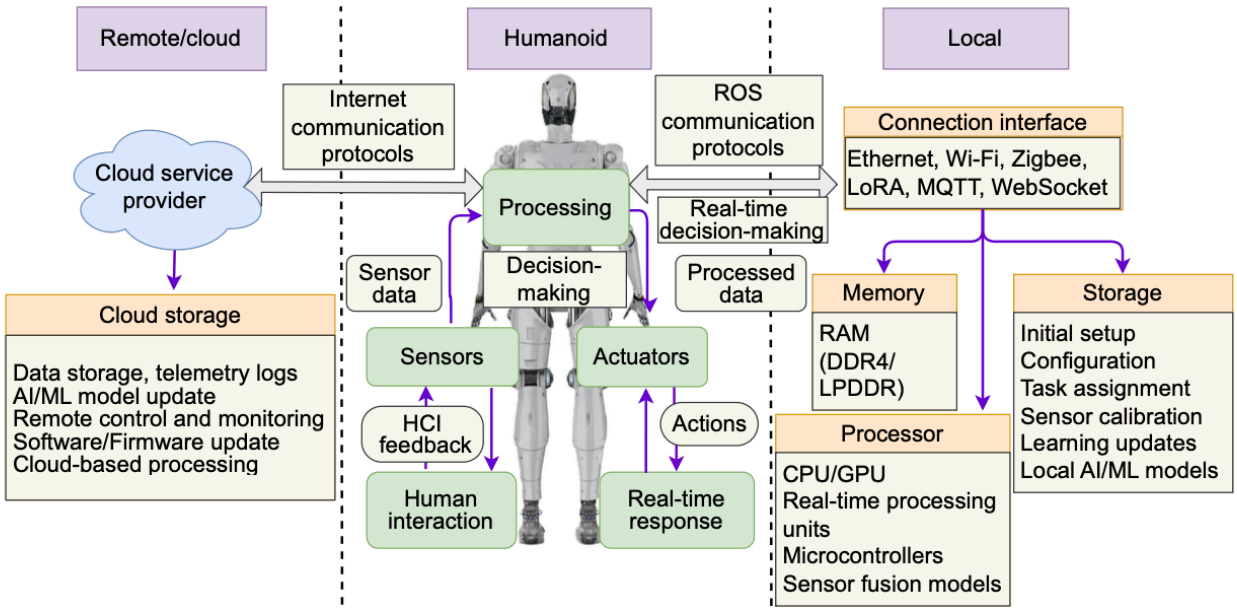

Figure 3: The humanoid in its cyber-physical ecosystem. On-board sensing, processing, and actuation form a local feedback loop for real-time control. Cloud-based components support asynchronous tasks such as learning, telemetry, and Over-The-Air updates. This view highlights that humanoids are nodes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

(Robey 2025)

This threat landscape is exacerbated by the deeply interconnected nature of humanoid architecture, where multilayered 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. 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. 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. Mitigating these varied, layered threats requires a comprehensive and specialized security approach. (Surve et al, 2025)

The security vulnerabilities extend even down to the operating system, Robot Operating System 2, which is an open source platform for control of robots. Not built from the ground up for security but for the special functions of robotic controllers.

Are Robot Vendors Securing their Products?

Given that HRs are to be deployed not just in infrastructure, industry but also the home, one would think security is a priority, however, there are many challenges to securing a complex system such as HRs, with many vulnerabilities:

> 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)

And it is not just these products that are susceptible even mega-cap companies are challenged by cybersecurity for HRs, suggesting funding defenses are not the problem:

> To better grasp how machine learning security helps to keep [Tesla] 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)

For Tesla Optimus one attack vector is shown to be possible theoretically:

> 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)

Humanoids Persist Problems in LLM Agents: Complex Insecurities

A recurring theme throughout this book has been the relationship of complexity to security vulnerabilities, the simple rule of: more complexity -> less security; is true in a certain sense, but can be mitigated, if the will and power are there to do such things, but it is easy to undermine security culture when the chief metric of success is capitalization or profits, and a constraint of having higher returns this day than the day before, a constant upward pressure. This is why the other theme of individual empowerment for cybersecurity is also in this book, ultimately you are the only one you can rely on to secure your systems, either you choose being a luddite or at the mercy of those more powerful then you if you choose not to do your best for your own security, though we all may end up being primitivists anyway. To explain the complexity Surve et al give a reasonable account of the chaotic butterfly effect AI security paradox:

> 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)

Security is different for isolated systems that are not interacting with other systems, as isolated nodes, secure one point and all points are secure. In the case of HRs we are dealing with many nodes and complex linkages between those nodes, the wiring is not straightforward. This is also true in Agentic AI.

Safety or Profits First?

So certainly vendors want safe systems for the consumers no matter how much it affects the price point, right? What vendors *have* disclosed about system security or what we can infer. Are vendors being accountable or even regulated for security?

Some of which would include vendor best-practices: “secure authentication, encrypted communication, and supply chain security are crucial” in the robotics domain (Rajashekaraiah, 2025)

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’”. (EvolveLabs, 2025)

- 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. (Olajide, 2025)

- But no auditable list of security is provided or commitments to security are provided in real terms as far as confirmation goes.

Unitree (China)

- 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.” (South China Morning Post, 2024)

- 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. (Mayoral-Vilches et al, 2025)

  Unitree Four critical failures 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 .
  • **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 autoreconnect 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.

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)

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 in commercial robotics platforms to date, much more mature than the industry average" (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**. (Mayoral-Vilches et al, 2025) 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).

- 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.

🔍 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, in a military context a malicious takeover of humanoid robots could turn them into an occupation army, theoretically speaking.

📋 Summary

- Many vendors are aware of the cybersecurity & physical safety risks for humanoid robots; they offer high-level commitments, but no public way to confirm that they are safe.

- 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.

- 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. See below discussion about Mandatory Cybersecurity Assurance Disclosures (MCAD)

Table of **major civilian-humanoid robot vendors** with publicly known cybersecurity disclosures, safety measures and incident reports.
(*Note: "disclosure" means what we found in open sources; many gaps remain.*)

HR Security Disclosures by Company (Top 10)

Vendor	Flagship Platform	Autonomy Model	Cybersecurity Posture (Public)	Safety / Control Measures	Public Incidents / Disclosures
Tesla	Optimus	High autonomy (vision-based, end-to-end learning)	Limited formal cybersecurity disclosure; relies on Tesla software security practices	Centralized compute, OTA updates, supervised deployment	No public Optimus incidents; Tesla vehicle cybersecurity incidents documented
Boston Dynamics	Atlas (research), Spot (commercial)	Semi-autonomous, task-bounded	Publishes security advisories for Spot; ROS2 hardening	E-stop, geofencing, teleop override	No major public cyber incidents
Figure AI	Figure 01	High autonomy, LLM-integrated	Minimal public cybersecurity disclosure (early stage)	Human-in-loop demos; supervised tasks	No public incidents
Agility Robotics	Digit	Semi-autonomous logistics	Participates in industrial safety standards; limited cyber detail	Redundant safety controllers, human supervision	No public incidents
Unitree Robotics	H1	Semi-autonomous	Sparse cybersecurity disclosures	Physical safety controls; demos supervised	No public incidents
UBTECH	Walker X	Semi-autonomous	Limited public cyber posture	Human-supervised operation	No public incidents
PAL Robotics	TALOS	Research / industrial	ROS-based; follows EU robotics safety norms	Kill-switches, bounded autonomy	No public incidents
SoftBank Robotics	Pepper	Low autonomy, cloud-connected	Publicly documented cloud architecture; privacy controls	Remote shutdown, limited actuation	Past Pepper cloud outages, no major cyber harm

Engine AI	PM01 / humanoid platforms	Semi- to high-autonomy (open platform)	**Open platform increases attack surface; no formal cyber standard published**	Research-focused; supervised demos	No public incidents
XPENG (Xpeng Robotics)	PX5 humanoid / embodied AI	High autonomy, AI-centric	**No public cybersecurity disclosures for humanoids**	Internal safety controls claimed; demos supervised	No public incidents

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, this is for architectural reasons, some for using LLMs in robotics as well.

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.

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.

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.

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.

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.

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. Yet, one can never fully anticipate how a consumer may use a robot, or how a military may need to use a robot in the midst of

kinetic actions on the battlefield. Again, we are dealing with regulations trailing innovation in the western nations.

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.

Countermeasures: Defense-in-Depth for Embodied AI

Managing humanoid robot risk requires treating them as **high-risk cyber-physical agents**, not appliances, or heavy industrial machines that are dangerous to operate that need fail safe mechanisms, just as any dual use technology should. Controls can come in both technical and structural modes, with the need to delineate between military operations robots and civilian use robots with different policies for both.

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.

Architectural Safeguards

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

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

As robotics is dual-use those robots intended for civilian use should not have the same functionalities that military robots would have, such as hardening— extra defenses against attack. Consider how having a military grade robot in a office or home could lead to weaponization in all spaces.

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.

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

Military Oversight

This raises the question of oversight for military use. Just as there are the Geneva Conventions on warfare, it would be necessary to add categories for autonomous robotics in warfare. The major actors are NATO, China, and Russia. In the western countries there is more public knowledge of oversight, in the PRC and Russia there are less public oversight mechanisms in place, the more troubling aspect is the lack of International Treaty law on the use of robotics in warfare. 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.” (Dohnal, 2025).

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’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.”
- There are standardisation efforts for unmanned / autonomous systems: e.g., STANAG 4671 covers unmanned aerial systems airworthiness for NATO.
- Studies of member-state strategies show awareness of autonomy and unmanned systems issues, including risk of cyber-attacks, need for governance and human oversight. (Gray et al, 2021)

.

⚠️ **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. (Marcinek et al, 2023)

- 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.

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.
-

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 the 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" language	Known or inferred gaps (especially mechanical/ hardware fail-safe)
North Atlantic Treaty Organization (NATO / Allies)	• "Autonomy Implementation Plan" (2022) – Allies commit to deploying autonomous systems consistent with the "Principles of Responsible Use". (NATO, 2022) • Paper: "Maintaining Appropriate Human Control in RIA Systems" – stresses human control and oversight. (Boardman, 2019)	• Emphasises "governability", "responsibility and accountability", "traceability" of systems. (NATO 2022) • Recognizes need for human-agent teams, oversight of autonomous decisions. (Boardman, 2019)	• Does *not* appear to mandate explicit **mechanical/hardware interlock fail-safe mechanisms** (e.g., physical keys, default locked actuators) in publicly accessible docs. • Less visibility on how systems should handle adversary interference (cyber/EM) in doctrine.
People's Republic of China (PLA / Chinese military robotics)	• Analysis: China heavily investing in robotics, autonomous/unmanned systems, swarms, robotics integrated into combined arms. (Höpflinger, , 2022)	• Public discussion focuses on using robotics to augment, reduce human manpower, and on battlefield efficiency. • Less publicly detailed normative language on "human in the loop" or oversight.	• Very limited transparent policy on mechanical fail-safe design or how adversary interference is handled. • Mechanisms for ensuring human override, hardware safing, or tamper resilience are not clearly specified in cited material.
Russian Federation (Russian military/ autonomy policy)	• "Russian Perceptions of Military AI, Automation, and Autonomy" – describes Russia's strategic priority for AI/robotics. (Nadibaidze, 2022) • RAND "Robotization of the Armed Forces" study – Russia sees robotization as asymmetric force option. RAND Corporation	• Emphasises automation/robotics to reduce manpower, enhance capability. • Some mention of automation and autonomy but human-control language weaker in open sources. Foreign Policy Research Institute	• Publicly accessible doctrine does *not* clearly articulate detailed oversight, human fail-safe, or mechanical interlock frameworks. • The risk of adversary takeover, cyber/EM interference, appears less addressed in visible Russian open material.

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 — *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 adversarial control

- 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.

Country / Regime	Key Policy/ Doctrine Reference	Human Control / Oversight Emphasis	Mention of Mechanical/ Hardware Fail-Safe or Interlock	Gaps & Comments
United States	CRS "U.S. Policy"	Yes – human judgment required. (US Congress, 2025)	Limited public detail on mechanical interlocks	Focus on human-in-loop but less on detailed hardware safeguards
NATO (Allied States)	NATO PA report 2023	Yes – governance, human control emphasised. (Weingarten, 2023)	Not much open detail visible	Normative framework exists, but engineering details missing
China	(open source limitation)	Public capacity emphasis, less published oversight detail	Very limited public hardware fail-safe discussion	Significant gap in open oversight docs
Russia	Military-automation analysis	Emphasis on autonomy / robotics capability. (Titriga, 2016)	Less visible public oversight/hardware detail	Higher risk of less accountability

Civilian Regulations

Proposal: Closing the Civilian Humanoid Robotics Cybersecurity Disclosure Gap

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:

Category	Required Disclosures	Example Metrics
System Updates	Frequency, authentication of OTA updates, rollback protection, verification of firmware signatures.	Update cadence, hash verification protocol, responsible disclosure timeline.
Access Control	Multi-factor authentication, password policy, physical service-port restrictions, default credential elimination.	List of privileged access interfaces and controls.
Network Isolation	Default network segmentation, firewall/whitelisting rules, external-service dependencies, data egress design.	Port and protocol exposure summary; remote telemetry endpoints.
Fail-Safe Defaults	Description of physical/electronic mechanisms to halt or limit actuation upon control loss or anomaly detection.	“Safe posture” state definition; manual override description.
Incident Response	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
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.

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

Boardman, M., et al. (2019). *An exploration of maintaining human control in AI-enabled systems and the challenges of achieving it*. NATO STO Meeting Proceedings (STO-MP-IST-178). https://publications.sto.nato.int

Boston Dynamics. (2023–2024). *Spot security and safety documentation*.

Congressional Research Service. (2025). *Defense primer: U.S. policy on lethal autonomous weapon systems*. https://www.congress.gov/crs-product/IF11150

Dohnal, J. (2025). *Legal aspects of the development of weapons systems with artificial intelligence in 2025*. https://arws.cz/news-at-arrows/legal-aspects-of-the-development-of-weapon-systems-with-artificial-intelligence-in-2025

Engine AI. (2024–2025). *Product demonstrations and humanoid platform announcements*.

European Commission. (2024). *Machinery Regulation and AI Act guidance*.

Evolvelab. (2025). *The dawn of the Tesla Bot: Revolutionizing automation*. https://evolverobot.in/optimus-the-tesla-bot

Gray, M., et al. (2021). *Artificial intelligence and autonomy in the military: An overview of NATO member states' strategies and deployment*. NATO CCDCOE. https://ccdcoe.org/library/publications/artificial-intelligence-and-autonomy-in-the-military-an-overview-of-nato-member-states-strategies-and-deployment/

Höpflinger, M. (2022). *Stand und Entwicklung militärischer Roboter*. stratos digital, No. 23. https://dam.gcsp.ch/files/doc/great-powers-military-robotics

Human Rights Watch. (2021). *Stopping killer robots: Country positions on banning fully autonomous weapons and retaining human control*.

INCIBE. (2024). *Security advisories affecting Unitree robotic platforms*. Spanish National Cybersecurity Institute.

International Committee of the Red Cross. (2021). *Autonomous systems and humanitarian risk*.

ISO/SAE. (2021). *ISO/SAE 21434: Road vehicles — Cybersecurity engineering*.

Madsen, T. (2025). *IEC 62443: A cybersecurity guide for industrial systems (Part 5)*.

Marcinek, K., et al. (2023). *Russia's asymmetric response to 21st-century strategic competition: Robotization of the armed forces*. RAND Corporation. https://www.rand.org/pubs/research_reports/RRA1233-5.html

Mayoral-Vilches, V. (2025). *Cybersecurity AI: Humanoid robots as attack vectors*. arXiv:2509.14139v1. https://github.com/aliasrobotics/cai

Nadibaidze, A. (2022). *Russian perceptions of military AI, automation, and autonomy*. Foreign Policy Research Institute. https://www.fpri.org/wp-content/uploads/2022/01/012622-russia-ai-.pdf

Naraine, R. (2025, April 1). *Hackers could unleash chaos through backdoor in China-made robot dogs*. *SecurityWeek*.

NATO. (2022). *Summary of NATO's autonomy implementation plan*. https://www.nato.int

NATO Parliamentary Assembly. (2023). *Robotics and autonomous systems report*. https://www.nato-pa.int

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*.

PAL Robotics. (2023). *TALOS humanoid technical overview*.

Rajashekaraiah, M. (2025). *Ensuring a secure future for robotics: The role of cybersecurity*. Analog Devices. https://www.analog.com/en/signals/thought-leadership/ensuring-a-secure-future-for-robotics.html

Robey, A., Ravichandran, Z., Kumar, V., Hassani, H., & Pappas, G. J. (2025). *Jailbreaking LLM-controlled robots*. arXiv:2508.17481v2.

SoftBank Robotics. (2020–2022). *Pepper architecture and cloud services documentation*.

South China Morning Post. (2024). *China's laws of robotics: Shanghai publishes first humanoid robot guidelines*. https://finance.yahoo.com/news/chinas-laws-robotics-shanghai-publishes-093000734.html

Stix, C., Hallensleben, A., Ortega, A., & Pistillo, M. (2025). *The loss of control playbook: Degrees, dynamics, and preparedness*. Apollo Research. arXiv:2511.15846.

Surve, P. P., Shabtai, A., & Elovici, Y. (2025). *SoK: Cybersecurity assessment of the humanoid ecosystem*. arXiv.

Tesla, Inc. (2023–2024). *AI Day and Optimus program materials*. https://www.tesla.com

The CDO Times. (2023). *Tesla's AI strategy and robotics ambitions*.

Titiriga, R. (2016). *Autonomy of military robots: Assessing the technical and legal ("jus in bello") thresholds*. *John Marshall Journal of Information Technology & Privacy Law*, 32, 57. https://repository.law.uic.edu/jitpl

Tri-City Voice. (2023). *China's robotics regulations and risk-warning requirements*.

UBTECH Robotics. (2023). *Walker X humanoid product documentation*.

Unitree Robotics. (2023–2024). *H1 humanoid demonstrations and OTA update disclosures*. https://www.unitree.com

US Congress. (2025). *Defense primer: U.S. policy on lethal autonomous weapon systems*. https://www.congress.gov

Weingarten, J. (2023). *Developing future capabilities: Robotics and autonomous systems*. NATO Parliamentary Assembly. https://www.nato-pa.int/document/2023-robotics-and-autonomous-systems-report-weingarten-034-stctts

XPENG Inc. (2024). *Embodied AI and humanoid robotics announcements*.