Physical AI Safety: A Definitional Framework

Mati Melchior
Independent Physical AI Safety Researcher
mati@physical-ai-safety.com

May 2026

Abstract

Physical AI is moving from research into industrial deployment at scale. The International Federation of Robotics reports more than four million industrial robots in operation globally, and recent national strategy documents identify Physical AI as a priority area for the coming decade [IFR 2025; IIA 2026]. Three discipline traditions inform the safety of such systems: AI Safety, Functional Safety, and Robot Safety. None of them, individually or collectively, addresses the failure modes that arise when machine-learning components drive physical actuators with assurance levels comparable to legacy automation. This paper introduces Physical AI Safety as a distinct discipline. It is the discipline of designing, verifying, and operating Physical AI systems such that the probability of physical harm is bounded to levels equivalent to or lower than those of comparable legacy automation under existing functional safety frameworks. The paper presents a four-layer reference model — the Physical AI Safety Stack — that spans AI decision, software safety monitoring, hardware safety constraint, and actuators. It enumerates a taxonomy of six failure modes specific to Physical AI: hallucinated commands, distributional shift, common-cause software failure, adversarial input, sensor-based deception, and emergent multi-agent failure. A research agenda of ten open problems follows. The category exists; the term does not yet. This paper proposes both, and invites the academic community, standards bodies, and regulators to adopt them.

Keywords: Physical AI Safety, AI safety, robot safety, functional safety, autonomous systems, hardware safety, certification.

---

1 Introduction

On April 17, 2024, the U.S. National Highway Traffic Safety Administration closed Engineering Analysis EA22002, having reviewed 956 crashes in which Tesla's Autopilot driver-assistance system was alleged to have been engaged. Of these, 13 crashes were fatal. The agency concluded that "driver disengagement, coupled with the Autopilot system design, is the key contributor to these crashes" [NHTSA 2024]. The case illustrates a recurring pattern: an AI system that controls physical actuators encounters operating conditions outside its design envelope, and the resulting failure

mode is contained neither by the AI nor by software-level safeguards alone.

The pattern is not isolated. It is structural. Physical AI systems — artificial intelligence systems whose outputs control physical actuators — are moving from research into deployment at scale. The safety frameworks under which they operate were designed for systems without learned components.

Three facts establish the urgency.

Scale. The International Federation of Robotics reports more than four million industrial robots in operation globally. South Korea operates 1,220 robots per ten thousand manufacturing workers, the highest density in the world [IFR 2025]. Venture investment in robotics reached approximately fourteen billion United States dollars in 2025 per Crunchbase 2025 venture data. Humanoid robotics, autonomous vehicles, and surgical robotics each show year-over-year growth in commercial deployment.

Capability. AI components inside Physical AI systems are increasing in capability. End-to-end learned policies now drive humanoid robots, autonomous vehicles, and warehouse logistics. Functions that were previously rule-based — perception, planning, control — are now learned. This shift transfers risk from deterministic, inspectable code to opaque statistical models.

Gap. No single existing safety framework addresses both the AI component and the physical effect. AI Safety addresses model behavior but not certification or hardware fault domains. Functional Safety (IEC 61508, ISO 13849) addresses electrical and electronic systems but provides no methodology for machine-learning components. Robot Safety (ISO 10218, ISO 13482) addresses mechanical hazards. The April 2026 Israel Innovation Authority national strategy report identifies the gap as a national priority [IIA 2026].

We propose Physical AI Safety as a distinct discipline. The remainder of this paper defines the term, distinguishes it from adjacent disciplines, presents a structural framework (the Physical AI Safety Stack), enumerates the failure modes the discipline must address, and proposes a research agenda.

---

2 Why a new discipline?

Four existing disciplines bear on the safety of Physical AI. Each covers part of the problem. None covers all of it. Table 1 summarizes the structure of the gap.

Table 1. Existing disciplines and where they fall short for Physical AI.

Discipline	Primary concern	Primary mechanism	Where it falls short for Physical AI
AI Alignment	Intent / values	Training, RLHF, oversight	Doesn’t address embodied failure modes; mostly cloud-focused

Discipline	Primary concern	Primary mechanism	Where it falls short for Physical AI
AI Safety (broader)	AI system behavior	Robustness, interpretability, oversight	Doesn't address hardware-level safety, certification
Functional Safety (IEC 61508)	Random + systematic faults in E/E/PE systems	SIL levels, redundancy, formal methods	Doesn't address AI components (no methodology for ML model safety)
Robot Safety (ISO 10218, ISO 13482)	Mechanical hazards from robots	Guards, separation, mechanical safety	Doesn't address AI-driven decision failures
Embodied AI (academic)	AI capability in physical contexts	Simulation, learning, planning	Performance-focused, not safety-focused
Physical AI Safety (this paper)	AI-driven physical harm	Stack-spanning hardware-software-AI methodology	—

AI Alignment. AI Alignment research addresses whether an AI system's objectives match human intent. Techniques include reinforcement learning from human feedback, constitutional methods, and interpretability research. The discipline has produced strong results for cloud-deployed language models. Its methods do not address the physical layer — actuator faults, sensor occlusion, hardware fault domains. A perfectly aligned model can still cause physical harm through hallucinated commands or sensor deception.

AI Safety (broader). AI Safety as a broader area covers robustness to distributional shift, adversarial robustness, oversight, and reward hacking [Amodei et al. 2016]. The Concrete Problems agenda identified five failure modes that remain central to the field. The agenda was framed against cloud and digital deployments. Hardware fault tolerance, certification methodology, and physical assurance fall outside its scope by design.

Functional Safety (IEC 61508). IEC 61508 and its sector derivatives provide a mature methodology for assuring electrical, electronic, and programmable electronic systems against random and systematic faults. Sector derivatives include ISO 13849 (machinery), ISO 26262 (road vehicles), IEC 62304 (medical device software), EN 50128 and EN 50657 (rail), DO-178C (aviation), and IEC 60880 (nuclear). The methodology is rigorous. Its assumptions, however, do not extend to machine-learning components: specification by training rather than by deterministic rules, opacity of decision processes, distributional shift at runtime. There is no SIL-level analogue for an ML model.

Robot Safety (ISO 10218, ISO 13482). Robot safety standards address mechanical hazards: pinch points, collision energy, separation distances, speed limits, emergency-stop pathways. ISO 13482 governs personal-care robots. ISO 10218 governs industrial robots. ISO/TS 15066 governs collaborative-robot operation. These

standards assume that the robot's behavior is specified by the system designer. They do not contemplate a robot whose behavior emerges from a learned model.

Embodied AI (academic). Embodied AI in the academic sense — sim-to-real transfer, learned manipulation, foundation models for robotics — is performance-oriented. Safety enters as a constraint within learning, not as a certifiable property of the deployed system. Where safety is addressed, it is typically through reward shaping or constraint satisfaction during training, not through structured assurance.

Physical AI Safety. Physical AI Safety is integrative, not adversarial. It does not replace any of the above disciplines. It draws on each. From AI Safety, it inherits the framing of distributional shift and adversarial robustness. From Functional Safety, it inherits the methodology of structured assurance and SIL-style decomposition. From Robot Safety, it inherits the hazard analysis of physical interaction. From Embodied AI, it inherits the technical vocabulary of learned policies. Its contribution is the synthesis: a discipline scoped to the joint failure space that none of these covers alone.

---

3 Definitions

This section presents the canonical definitions of the discipline. The exact wording is offered for adoption by other researchers, standards bodies, and regulators.

3.1 Physical AI

> **Physical AI** denotes any artificial intelligence system whose outputs directly or indirectly control physical actuators that affect the physical world. Examples include humanoid robots, autonomous vehicles, surgical robots, drones, autonomous mobile robots, autonomous infrastructure (smart grids, autonomous water systems), and AI-controlled industrial process equipment. The defining characteristic is the closed loop between AI decision-making and physical effect on humans, property, or environment — not merely sensing or recommendation.

Excludes: AI systems whose outputs are advisory, informational, or limited to digital outputs (chatbots, recommendation engines, classifiers without actuator coupling).

The decisive criterion is actuator coupling. A perception system that ranks medical-imaging findings is not Physical AI; the same perception system embedded in a surgical robot whose arm acts on the ranking is Physical AI. The boundary is not always crisp. Borderline cases are addressed in Table 2.

Figure 3 surveys nine domains in which Physical AI is currently deployed or under active development.

Figure 3. Domains of Physical AI deployment.

Table 2. Examples of systems that are or are not Physical AI.

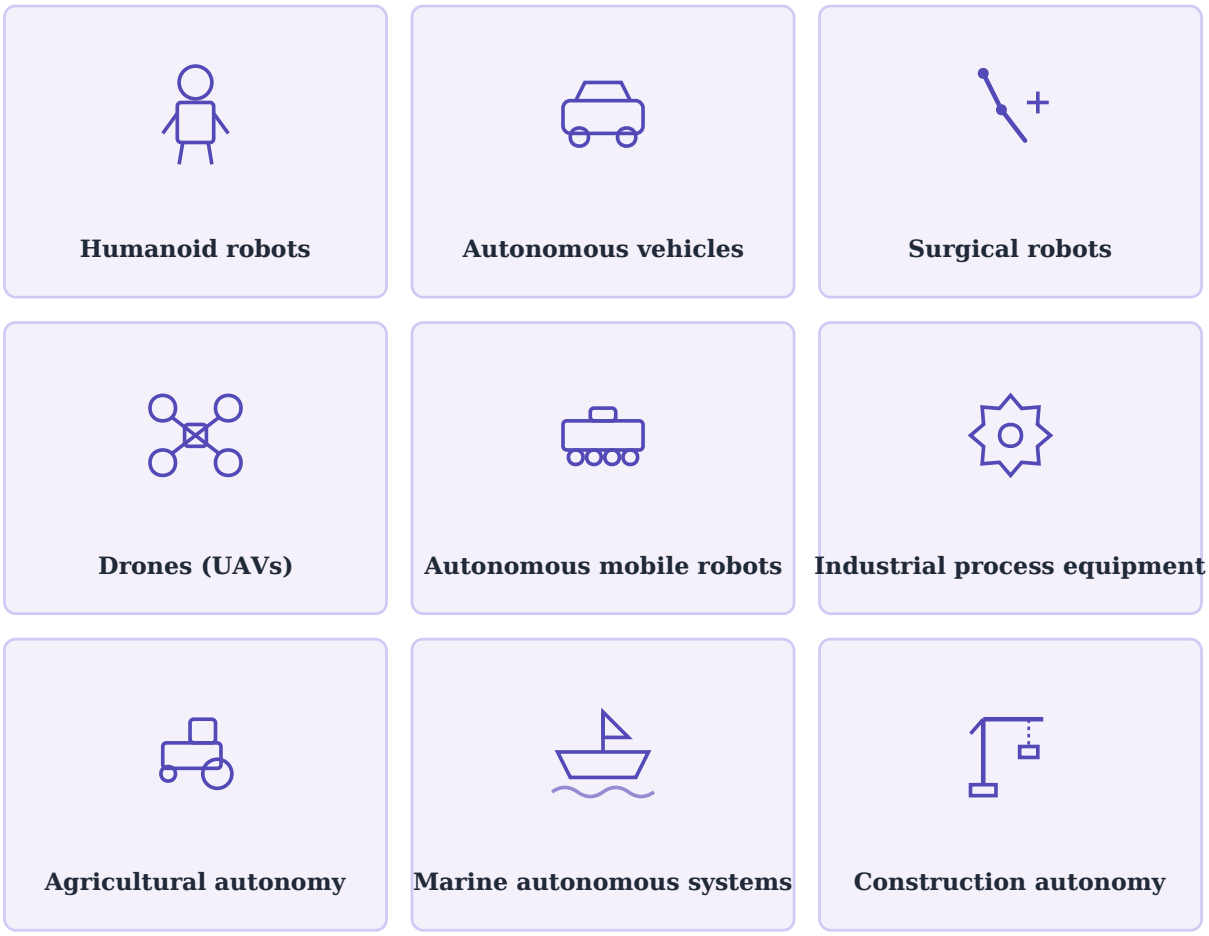

Figure 1: Three-by-three grid of nine Physical AI deployment domains, each shown with a simple geometric icon and label. Top row: humanoid robots, autonomous vehicles, surgical robots. Middle row: drones (UAVs), autonomous mobile robots, industrial process equipment. Bottom row: agricultural autonomy, marine autonomous systems, construction autonomy.

System	Physical AI?	Why
Tesla Autopilot	Yes	AI controls actuators (steering, braking, throttle)
Surgical robot (autonomous mode)	Yes	AI controls actuators (instruments)
Surgical robot (teleoperated)	Borderline	AI assists but human commands actuators directly
ChatGPT	No	No actuator coupling
Recommender system	No	Output is informational
ABB IRB cobot (programmed paths only)	No	No AI in the control loop
ABB IRB cobot with vision-guided AI picking	Yes	AI controls actuator decisions
Smart thermostat	Borderline	Actuator coupling, but trivially low harm potential
Autonomous mining equipment	Yes	High harm potential, AI control loop

3.2 Physical AI Safety

> **Physical AI Safety** is the discipline of designing, verifying, and operating Physical AI systems such that the probability of physical harm to humans, property, and environment is bounded to levels equivalent to or lower than those required of comparable legacy automation systems under existing functional safety frameworks (e.g., IEC 61508, ISO 13849, ISO 13482, ISO 26262).

Distinguishing characteristics:

- (i) Addresses failure modes that arise from AI components specifically (hallucination, distributional shift, adversarial input, opaque reasoning)
- (ii) Requires assurance levels comparable to legacy automation, not merely "reasonable effort"
- (iii) Operates across the full system stack — algorithmic, software, hardware, physical
- (iv) Distinct from AI Alignment (which addresses intent), Functional Safety (which addresses non-AI components), and Robot Safety (which addresses mechanical hazards)

Two phrases bear emphasis. *"Equivalent to or lower than"* is a calibrated claim: Physical AI is not held to a stricter standard than legacy automation, but it is held to the same standard. *"Existing functional safety frameworks"* names the calibration target — IEC 61508 and its derivatives, which define what *bounded probability of harm*

means in operational terms (failure rates per hour, SIL levels, dangerous-failure budgets). Physical AI Safety adopts the calibration; it does not redefine it.

Characteristic (iii) is the central design constraint. A discipline that operates only at the model layer cannot bound physical harm. A discipline that operates only at the actuator layer cannot address AI-specific failure modes. The full stack is required.

3.3 The Physical AI Safety Stack

The **Physical AI Safety Stack** is a four-layer reference model:

Layer 1 (L1) — AI / decision: the AI components that produce intent, plans, and commands. ML models, planners, policies.

Layer 2 (L2) — Software safety monitoring: software that observes Layer 1 outputs and applies safety constraints, watchdogs, and runtime assurance.

Layer 3 (L3) — Hardware safety constraint: dedicated hardware that enforces safety limits independently of L1 and L2, with separation of fault domains (clock, power, memory, supply chain).

Layer 4 (L4) — Actuators and physical interface: the motors, valves, end-effectors, vehicle controls — the boundary at which decisions become physical effects.

Failure modes propagate downward; safety constraints propagate upward. The premise of Physical AI Safety is that Layer 3 cannot be eliminated by improving Layers 1 and 2.

A detailed treatment of the Stack follows in Section 4.

3.4 Failure modes specific to Physical AI

The discipline addresses six classes of failure mode. Detailed treatment follows in Section 5.

F1 — Hallucinated commands: AI produces structurally valid commands that are semantically incorrect for the current physical context (e.g., a humanoid commanded to grasp empty air, an autonomous vehicle planning a path through a wall).

F2 — Distributional shift: AI encounters physical conditions outside training distribution; behavior is unconstrained.

F3 — Common-cause software failure: redundant software channels share fate due to common OS, firmware, design assumptions; correlated failure invalidates redundancy claim.

F4 — Adversarial input: sensor or input streams crafted to induce unsafe behavior (visual, acoustic, electromagnetic).

F5 — Sensor-based deception: the AI's perception of the world diverges from physical reality due to sensor failure, occlusion, or spoofing.

F6 — Emergent multi-agent failure: in fleets, individually-safe agents produce collectively unsafe behavior (cascading failures, resource contention, collision dynamics).

3.5 Physical AI Safety Assurance

Physical AI Safety Assurance is the body of evidence demonstrating that a Physical AI system meets its safety requirements with quantified confidence. It includes hazard analysis, safety case documentation, third-party certification, hardware-level verification, and operational monitoring data. Distinct from "safety claims" (assertions without evidence) and "safety testing" (a subset of assurance).

The distinction between *claim*, *testing*, and *assurance* is operational. A safety claim is an assertion. A safety test is one form of evidence. Assurance is the structured argument that the test evidence, the design evidence, and the operational evidence collectively justify the claim. Functional safety practice treats this distinction as foundational; Physical AI Safety inherits it.

3.6 Other terms used in this paper

This paper introduces four additional terms that recur across the series. The first two are used in this paper. The latter two are introduced here only and developed in forthcoming companion papers.

- **Physical AI Safety Gap** — quantified distance between Physical AI capability investment and Physical AI Safety infrastructure investment in a market.
- **Safety-cert effort visibility** — degree to which a Physical AI organization's safety certification activities are publicly traceable.
- **PAS Level** — a maturity level (1–5) under the Physical AI Safety Maturity Model. The model is the subject of a forthcoming companion paper. This paper introduces the term only.
- **Regulatory Convergence Window** — the 2025–2027 period during which EU, US, Japan, and Israel regulatory regimes for Physical AI converge on third-party-certifiable hardware safety requirements. The argument is presented in a forthcoming companion paper. This paper introduces the term only.

---

4 The Physical AI Safety Stack

Figure 1 presents the four-layer reference model.

Figure 1. The Physical AI Safety Stack. *Top-down:* intent and commands propagate from L1 through L2 and L3 to L4. *Bottom-up:* safety constraints and feedback propagate from L4 toward L1. *Annotation:* fault propagation downward; assurance must propagate upward.

Layer 1 — AI / decision. The topmost layer contains the components that produce intent. Examples include end-to-end policy networks, model-predictive plan-

Figure 2: Vertical four-layer stack diagram. L1 AI decision (lightest purple) at top, L2 software safety monitoring, L3 hardware safety constraint (darkest purple), L4 actuators (gray) at bottom. Downward arrow on the left labeled “commands (intent)”. Upward arrow on the right labeled “assurance (safety constraints)”. Annotation below: fault propagation downward; assurance must propagate upward.

ners, learned controllers, perception–prediction–planning pipelines, and foundation models adapted for embodied tasks. Layer 1 is the source of the system's capability and the source of its AI-specific failure modes. Hallucination, distributional shift, adversarial vulnerability, and opaque reasoning all originate here.

Layer 2 — Software safety monitoring. Layer 2 observes Layer 1 outputs and applies constraints in software. Examples include runtime assurance shields, watchdog timers, output bound-checkers, plausibility filters, formally verified safety filters, and safe-set projection. Layer 2 catches a meaningful fraction of Layer 1 errors. Its limit is structural: it shares the execution environment with Layer 1 (operating system, hypervisor, processor, memory, power supply, firmware). A common-cause failure in that environment fails both layers simultaneously. F3 — common-cause software failure — names this failure mode.

Layer 3 — Hardware safety constraint. Layer 3 enforces safety limits in dedicated hardware that does not share fault domains with Layers 1 and 2. *Independent fault domains* means independent clock source, independent power supply rail, independent memory, separated supply chain, and distinct design lineage. Examples of constraints enforced at Layer 3 include velocity bounds, force and torque bounds, geometric limits (no-go zones), interlock state machines, and emergency-stop pathways. Layer 3 is the layer at which assurance can be made independent of the AI system's correctness.

Layer 4 — Actuators and physical interface. Layer 4 is the boundary at which command becomes effect. Motors deliver torque. Valves change flow. End-effectors apply force. Vehicle controls move steering, throttle, and brake. Below Layer 4 is the world: humans, property, environment. Layer 4 is where harm becomes physical.

The Stack is read in two directions. *Downward:* intent flows from L1 through L2 and L3 to L4. *Upward:* assurance must be argued from L4 back to L1. Failure modes propagate downward; safety constraints propagate upward. The premise of Physical AI Safety is that Layer 3 cannot be eliminated by improving Layers 1 and 2; this premise is argued in detail in our forthcoming companion paper on the software safety ceiling.

---

5 Taxonomy of failure modes specific to Physical AI

This section expands the F1–F6 taxonomy with concrete examples and coverage analysis against adjacent disciplines.

F1 — Hallucinated commands. AI produces structurally valid commands that are semantically incorrect for the current physical context. *Hypothetical example:* a humanoid robot operating in a kitchen receives the instruction "place the cup on the table" and executes a grasp on empty air because its world model placed the cup at a phantom location. The command is well-formed: position, velocity, and gripper-state are within actuator limits. It is wrong only because the world is not as the model represents it. Functional Safety frameworks check actuator limits, not world-model fidelity. AI Alignment addresses intent, not perception. Hallucinated commands are a Physical-AI-specific class.

F2 — Distributional shift. AI encounters physical conditions outside its training distribution; behavior is unconstrained. *Hypothetical example:* an autonomous vehicle trained on dry-pavement and rain conditions encounters a section of road covered in spilled motor oil. The friction model is outside the training distribution. The control policy issues commanded torque values that would be safe on the training distribution and unsafe on the actual road. Functional Safety covers component faults, not statistical out-of-distribution events. Robot Safety has no concept of training distribution.

F3 — Common-cause software failure. Redundant software channels share fate due to common operating system, firmware, design assumptions; correlated failure invalidates the redundancy claim. *Hypothetical example:* a robotic system runs two redundant safety monitors on the same operating system. A kernel bug — race condition, memory corruption, scheduler stall — affects both monitors simultaneously. The redundancy claim assumed independent failure modes; the shared platform invalidates the assumption. The seminal Knight–Leveson study [1986] reported that even N-version programs written independently exhibit correlated failures; subsequent system-safety literature has applied the same independence-failure logic to shared-platform redundancy [Leveson 1995]. IEC 61508 acknowledges common-cause failure (the β-coefficient) but provides no methodology for ML-based channels.

F4 — Adversarial input. Sensor or input streams crafted to induce unsafe behavior (visual, acoustic, electromagnetic). *Example:* an adversarial sticker placed on a stop sign causes a vision system to classify it as a 45-mph speed-limit sign. The phenomenon was first reported in laboratory conditions [Eykholt et al. 2018] and has since been extended to physical-world settings. Functional Safety has no model for adversarial input — its threat model is faults, not adversaries. AI Safety addresses the failure mode in the cloud setting; the embodied case adds physical-world realizability constraints.

F5 — Sensor-based deception. The AI's perception of the world diverges from physical reality due to sensor failure, occlusion, or spoofing. *Hypothetical example:* a LiDAR-equipped autonomous vehicle is targeted by a coordinated laser-injection attack that creates phantom obstacles. The sensor returns syntactically valid range data corresponding to objects that do not exist. The AI plans around the phantom. Robot Safety covers physical sensor faults (a stuck sensor) but not active deception of correctly-functioning sensors. Functional Safety addresses sensor fault detection, not adversarial spoofing.

F6 — Emergent multi-agent failure. In fleets, individually-safe agents produce collectively unsafe behavior (cascading failures, resource contention, collision dynamics). *Hypothetical example:* a fleet of warehouse mobile robots each correctly follows its individually-verified routing policy. A high-density configuration of orders produces a deadlock at a corridor intersection that no single robot's policy contemplated. The deadlock cascades into collision when timeouts misalign. Single-robot certification does not compose. No existing framework treats fleet-level safety as a primary concern.

Figure 2 summarizes coverage of the six failure modes by the adjacent disciplines.

Figure 2. Failure mode coverage matrix.

Failure mode	AI Alignment	Functional Safety	Robot Safety	Embodied AI	Physical AI Safety
F1 — Hallucinated commands	Partial	No	No	Partial	Yes
F2 — Distributional shift	Partial	No	No	Partial	Yes
F3 — Common-cause SW failure	No	Yes	No	No	Yes
F4 — Adversarial input	Partial	No	No	Partial	Yes
F5 — Sensor deception	No	Partial	Partial	No	Yes
F6 — Emergent multi-agent	No	No	No	Partial	Yes

Yes Partial No

Figure 3: Heatmap matrix of six Physical AI failure modes (F1 hallucinated commands, F2 distributional shift, F3 common-cause software failure, F4 adversarial input, F5 sensor deception, F6 emergent multi-agent failure) versus five disciplines (AI Alignment, Functional Safety, Robot Safety, Embodied AI, Physical AI Safety). Green cells indicate Yes coverage, amber cells indicate Partial coverage, gray cells indicate No coverage. The Physical AI Safety column is fully green across all six rows. No other column has more than one Yes cell.

Three observations follow from Figure 2. First, no adjacent discipline provides "Yes" coverage for more than one failure mode. Second, Functional Safety is the only discipline with established "Yes" coverage of F3, but it provides that coverage for non-AI channels only; the methodology requires adaptation to ML channels. Third, "Partial" coverage is not equivalent to *certifiable* coverage. A discipline that addresses a failure mode in research literature without a path to third-party certification leaves the operational gap unclosed.

---

6 A Physical AI Safety research agenda

The discipline is young. The following ten problems define the near-term research agenda. Each is open in the sense that no current methodology delivers a certifiable answer.

1. Quantifying Physical AI Safety Assurance. Functional safety expresses assurance through Safety Integrity Levels (SIL), defined in terms of dangerous-failure rates per hour. ML components do not have a comparable rate metric. What metric replaces SIL when the system contains ML components? Candidate framings include equivalence to a SIL-rated comparator, statistical performance bounds with confidence intervals, and structured assurance cases with explicit ML-specific evidence categories. None has been validated against the operational standard.

2. Hardware–software safety co-design for ML systems. Given the four-layer Stack, how should safety responsibilities be partitioned across layers? Which classes of constraint admit purely-software enforcement, and which require hardware enforcement? The answer is not uniform across actuator types, harm-severity classes, or operational environments.

3. Distributional-shift detection at the hardware safety boundary. Layer 3 must, in some configurations, detect that Layer 1 has gone outside its training distribution. This is non-trivial: Layer 3 by design does not have access to Layer 1's internal state. What architectural patterns enable distributional-shift detection at the Layer 3 boundary using only externally observable signals?

4. Common-cause failure analysis for ML+software hybrid channels. IEC 61508-6 quantifies common-cause failure through the β-coefficient. The standard methodology does not contemplate channels that share a training distribution, share foundation-model lineage, or share data-augmentation pipelines. How is β computed for ML+software hybrid channels?

5. Certification methodology for opaque AI components. Third-party certification bodies currently lack a methodology to evaluate the internals of an ML component they cannot fully inspect. What is the analogue of source-code review and structural coverage for an opaque model?

6. Physical AI Safety for fleets. Single-robot safety properties do not compose. A fleet of individually-safe robots is not collectively safe. Open: fleet-level safety frameworks that admit decomposition into single-agent properties plus interaction invariants.

7. Adversarial robustness verification. The literature on adversarial robustness has produced empirical defenses and a smaller set of certified bounds. The certified bounds rarely scale to the input dimensionality of physical-world sensors. How is robustness to *physical-world realizable* adversarial inputs verified?

8. Human–Physical AI interaction safety. When humans and Physical AI share workspaces dynamically — collaborative robots, shared roads, hospital corridors — the failure space includes joint failure modes that single-agent analysis does not surface. Specific failure modes and their mitigation remain open.

9. The verification gap. Current formal-verification tools handle subsystems of Physical AI: controllers, simple perception modules, isolated planners. None handles the full stack including learned components at production scale. Closing this gap requires methodology, not only tooling.

10. Evaluation benchmarks. No standardized benchmarks exist for Physical AI Safety. Benchmark design itself is an open problem: a benchmark must reflect operational risk, not test-set performance. The community needs benchmarks that admit certifiable claims.

These problems are not exhaustive. They are the tractable starting set.

---

7 Related work

This section surveys the four research traditions on which Physical AI Safety draws.

AI Safety literature. The modern AI Safety research agenda traces to Russell, Dewey, and Tegmark [2015] and Amodei et al. [2016]. The Concrete Problems agenda enumerated five failure modes — negative side effects, reward hacking, scalable oversight, safe exploration, and robustness to distributional shift — and proposed research directions for each. Subsequent work has expanded each direction into a sub-literature. The Anthropic AI Safety Levels (ASL) framework, modeled on the biosafety level (BSL) precedent, defines maturity tiers for model risk classes. The literature is predominantly model-centric and deployment-agnostic; the embodied case has received less direct treatment.

Functional Safety standards. IEC 61508 is the foundational functional safety standard for electrical, electronic, and programmable electronic systems. Sector derivatives include ISO 26262 (road vehicles), EN 50128 and EN 50657 (rail), IEC 62304 (medical device software), DO-178C (aviation software), and IEC 60880 (nuclear safety software). Each defines a structured assurance methodology with quantitative integrity levels, hazard analysis procedures, and component-decomposition rules. The methodology is mature for non-AI components. Application to ML components is an active open problem; ISO/PAS 8800 [2024] and the ISO/IEC 42001 family represent recent steps in that direction.

Robot Safety standards. ISO 10218-1 and -2 [2025] govern industrial robot safety. ISO 13482 [2014] governs personal-care robot safety. ISO/TS 15066 [2016] governs collaborative-robot operation. The European Machinery Regulation (EU) 2023/1230 and the European AI Act (EU) 2024/1689 together create the regulatory frame within

which European Physical AI deployments must be certified. United States frameworks include the NIST AI Risk Management Framework and sector-specific regulation by NHTSA, FAA, and FDA. The Hashemi-Petroodi et al. [2024] analysis of OSHA-recorded robot accidents documents the empirical failure landscape.

Embodied AI literature. Embodied AI as a research field encompasses sim-to-real transfer, learned manipulation, navigation, and end-to-end learned control. Foundation models adapted for robotics are a recent extension. The literature is performance-oriented; safety enters as a constraint within learning, not as a certifiable property of the deployed system. Hadfield-Menell et al. [2016] on cooperative inverse reinforcement learning is one of the few research lines explicitly tying embodied performance to alignment-flavored safety guarantees. Lipton [2018] on the disambiguation of "interpretability" provides a useful methodological precedent for this paper's disambiguation move.

Physical AI Safety draws from each of these traditions and reorganizes their contributions around a shared structural framework — the four-layer Stack — and a shared failure taxonomy.

---

8 Conclusion

Physical AI Safety is the discipline of designing, verifying, and operating Physical AI systems such that the probability of physical harm is bounded to levels equivalent to or lower than those of comparable legacy automation. It addresses the failure modes that arise when machine-learning components drive physical actuators. It is distinct from AI Alignment, AI Safety in the broad sense, Functional Safety, Robot Safety, and Embodied AI; it integrates contributions from each.

The discipline rests on two structures. The four-layer Physical AI Safety Stack — L1 AI decision, L2 software safety monitoring, L3 hardware safety constraint, L4 actuators — partitions the problem along the path from intent to physical effect. The six-class failure taxonomy — F1 hallucinated commands, F2 distributional shift, F3 common-cause software failure, F4 adversarial input, F5 sensor-based deception, F6 emergent multi-agent failure — names the failure modes the discipline must address. Together they organize the research agenda into ten near-term problems, none of which is currently solved.

The window is narrow. Industrial Physical AI deployment is scaling now. Certification methodology is not yet ready. The community — academic researchers, standards bodies, certification authorities, regulators, and industry — needs a shared vocabulary, a shared structural framework, and a shared agenda. This paper proposes the first two and contributes a starting set of problems for the third. The next decade of robotics will be the decade of Physical AI Safety.

References

[1] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Con-

crete Problems in AI Safety," *arXiv preprint arXiv:1606.06565*, 2016.

[2] S. Russell, D. Dewey, and M. Tegmark, "Research Priorities for Robust and Beneficial Artificial Intelligence," *AI Magazine*, vol. 36, no. 4, pp. 105–114, 2015.

[3] International Electrotechnical Commission, "IEC 61508:2010 — Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems," Parts 1–7, 2010.

[4] International Organization for Standardization, "ISO 10218-1:2025 — Robotics — Safety Requirements — Part 1: Industrial Robots," 2025.

[5] International Organization for Standardization, "ISO 13482:2014 — Robots and Robotic Devices — Safety Requirements for Personal Care Robots," 2014.

[6] International Organization for Standardization, "ISO 13849-1:2023 — Safety of Machinery — Safety-Related Parts of Control Systems — Part 1: General Principles for Design," 2023.

[7] European Union, "Regulation (EU) 2023/1230 of the European Parliament and of the Council on Machinery (Machinery Regulation)," *Official Journal of the European Union*, 2023.

[8] European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (AI Act)," *Official Journal of the European Union*, 2024.

[9] N. G. Leveson, *Safeware: System Safety and Computers*. Addison-Wesley, 1995.

[10] J. C. Knight and N. G. Leveson, "An Experimental Evaluation of the Assumption of Independence in Multi-Version Programming," *IEEE Transactions on Software Engineering*, vol. SE-12, no. 1, pp. 96–109, 1986.

[11] International Federation of Robotics, "World Robotics 2025 — Industrial Robots," 2025.

[12] Israel Innovation Authority, "National Artificial Intelligence Strategy Report," April 2026.

[13] Anthropic, "Anthropic's Responsible Scaling Policy" (AI Safety Levels framework), 2023.

[14] International Electrotechnical Commission, "IEC 60880 — Nuclear Power Plants — Instrumentation and Control Systems Important to Safety — Software Aspects for Computer-Based Systems Performing Category A Functions," 2006.

[15] RTCA, "DO-178C — Software Considerations in Airborne Systems and Equipment Certification," 2011.

[16] CENELEC, "EN 50128:2011/A2:2020 — Railway Applications — Communication, Signalling and Processing Systems — Software for Railway Control and Protection Systems," 2020.

[17] S. E. Hashemi-Petroodi et al., "Robot-Related Occupational Accidents: An Analysis of OSHA Records," *ScienceDirect*, 2024.

[18] K. Eykholt et al., “Robust Physical-World Attacks on Deep Learning Visual Classification,” in *Proc. CVPR*, 2018.

[19] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan, “Cooperative Inverse Reinforcement Learning,” in *Proc. NeurIPS*, 2016.

[20] International Organization for Standardization, “ISO/PAS 8800 — Road Vehicles — Safety and Artificial Intelligence,” 2024.

[21] Z. C. Lipton, “The Mythos of Model Interpretability,” *Communications of the ACM*, vol. 61, no. 10, pp. 36–43, 2018.

[22] International Organization for Standardization, “ISO 26262:2018 — Road Vehicles — Functional Safety,” 2018.

[23] International Organization for Standardization, “ISO/TS 15066:2016 — Robots and Robotic Devices — Collaborative Robots,” 2016.

[24] U.S. National Highway Traffic Safety Administration, “Standing General Order on Crash Reporting” (incident data on automated and semi-automated vehicle systems).

[25] U.S. National Highway Traffic Safety Administration, Office of Defects Investigation, “Investigation Closing Report — Engineering Analysis EA22002 (Tesla Autopilot),” April 25, 2024. [Online]. Available: https://static.nhtsa.gov/odi/inv/2022/INCR-EA22002-14496.pdf

---

Mati Melchior · Independent Physical AI Safety Researcher mati@physical-ai-safety.com · physical-ai-safety.com P1 of an 8-paper series on Physical AI Safety.