Towards a Physical AI Safety Certification Framework: A Synthesis and Proposal

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

Version 1.0 · May 2026

Abstract

The deployment of artificial intelligence in physical systems — robots, autonomous vehicles, surgical platforms, and other actuator-driven machines — has outpaced the safety standards meant to govern it. This paper synthesizes seven prior contributions into a proposed certification framework for Physical AI: the Physical AI Safety Certification Framework, or PAS-CF.

The argument proceeds from convergence. Three factors point to the same gap. Regulatory pressure is the first: the EU Machinery Regulation, the EU AI Act, and parallel work in other jurisdictions. The second is technical: common-cause failure analysis in machine-learning-bearing safety channels. The third is empirical: a maturity assessment across the industry. Each factor independently surfaces the same finding — existing functional safety standards do not yet contain AI-specific evaluation criteria.

PAS-CF proposes four such criteria. The first covers AI behavior monitoring under distributional shift. The second covers the presence and integrity of a hardware-layer safety mechanism with separation of fault domains. The third covers common-cause failure analysis with explicit β-coefficient reporting. The fourth covers the audit trail and incident response capability of the system in operation.

The framework augments rather than replaces existing standards. It maps onto IEC 61508, ISO 13849, ISO 13482, ISO 10218, and the emerging ISO/IEC TR 5469. We illustrate application across three example architectures: an industrial cobot with vision-guided picking, an autonomous mobile robot for warehouse logistics, and a surgical robot operating in autonomous mode. The implementation roadmap proceeds in three horizons: voluntary self-assessment (12 months), formal proposal to standards bodies (3 years), and a derived ISO/IEC standard with mandatory third-party certification for high-risk Physical AI (5 plus years).

PAS-CF is offered as a starting point for standards-body deliberation. The paper closes with an open invitation to ISO TC299, IEC TC65, certifying bodies, regulators, industry, and researchers to engage with and improve the framework.

Keywords: Physical AI Safety, certification, IEC 61508, ISO 13849, ISO 13482, ISO 10218, standards, third-party assessment, framework

1 Introduction and Context

The Regulatory Convergence Window for Physical AI Safety is closing. Regulation (EU) 2023/1230 (the EU Machinery Regulation) takes effect in January 2027. By that date, certifiable Physical AI safety frameworks must exist, or industry will rely on ad-hoc approaches that may not survive third-party audit. The EU AI Act (Regulation (EU) 2024/1689) runs on a parallel timeline that adds compliance pressure. Similar trajectories appear in the United States, Japan, and Israel.

A four-jurisdiction convergence pattern is now established (the survey of regulatory trajectories is the subject of a forthcoming companion paper). The pattern is unambiguous: AI-bearing safety-relevant systems will require third-party certification within the decade. The question is not whether but how.

This paper is the eighth in a series that has established the disciplinary foundations of Physical AI Safety. Where it draws on the prior contributions, it does so by topic, with each contribution summarized below.

Forthcoming companion paper on…	Contribution
Definitional framework	The term *Physical AI*, the four-layer Physical AI Safety Stack, and the F1–F6 failure-mode taxonomy
The Software Safety Ceiling	Software-only architectures cannot evaluate their own decision layer
Reference architecture	The hardware safety co-processor and three architectural patterns (in-path, side-monitor, gateway)
Maturity model	The PAS-MM five-level model, modeled on functional-safety SIL/PL conventions
Empirical baseline	The Physical AI Safety Gap surfaced by an industry survey
Common-cause failure analysis	Methodology for β-coefficient estimation in ML-bearing channels
Regulatory mapping	A four-jurisdiction analysis of the convergence window

These seven companion contributions describe the discipline. They do not, individually or jointly, constitute a certification framework.

This paper proposes such a framework. The Physical AI Safety Certification Framework — PAS-CF — integrates the prior work into a structure that can be evaluated against, audited against, and (over time) certified against. PAS-CF is not a final standard. It is a starting point for standards-body deliberation.

Framing. Two framings of the contribution are possible. The first is rejection: the existing safety-standards corpus does not address Physical AI, and a new framework

must replace it. The second is augmentation: the existing corpus addresses most of what Physical AI requires, and a new framework adds the AI-specific portion. This paper takes the second framing. The reasons are practical (the existing corpus represents decades of accumulated safety engineering experience) and structural (the gaps are localized to AI-specific concerns, not pervasive). The framing matters because it determines who PAS-CF is for: not for those who would discard IEC 61508 and ISO 13849, but for those who would extend them.

Audience. The audience for this paper includes ISO TC299 and IEC TC65 working group delegates, who may consider PAS-CF as input for new work; certifying bodies (TÜV SÜD, TÜV Rheinland, SGS, Pilz, and Notified Bodies under EU 2023/1230), who may pilot PAS-CF as augmentation to existing certification; government agencies considering Physical AI regulation; compliance officers at Physical AI vendors; insurance underwriters; and researchers. The framework is intended to be readable by all of these audiences.

Roadmap. Section 2 recaps the four-layer Physical AI Safety Stack and draws out its certification implications. Section 3 conducts gap analysis on the major existing standards. Section 4 introduces the four AI-specific evaluation criteria of PAS-CF. Section 5 maps PAS-CF onto existing standards. Section 6 applies PAS-CF to three example architectures. Section 7 sketches an implementation roadmap. Sections 8 and 9 cover limitations, future work, and open research questions. Section 10 concludes with a call to action.

---

2 The Four-Layer Stack and Certification Implications

2.1 The Physical AI Safety Stack

The Physical AI Safety Stack, introduced in our forthcoming companion paper on the definitional framework, decomposes a Physical AI system into four layers. Figure 1 shows the structure.

Figure 1. The Physical AI Safety Stack (redrawn from the forthcoming definitional paper). The four layers separate decision, software-safety monitoring, hardware-level constraint enforcement, and physical actuation. Each layer has a distinct role and distinct certification needs.

The Stack is descriptive, not prescriptive. It exists in any Physical AI system whether or not its designers were aware of it. The certification question is: to what assurance level does each layer perform its role, and how is the interaction between layers verified?

Our forthcoming companion paper on the software safety ceiling argues that a system with only L1 and L2 — ML decision layer plus software monitor — cannot reach high assurance levels for safety-critical operation. The reason is structural: a software safety monitor that runs on the same compute substrate as the ML system it monitors shares fault modes with what it monitors. The companion paper terms this *fate-sharing*. The monitor cannot evaluate the monitored from outside.

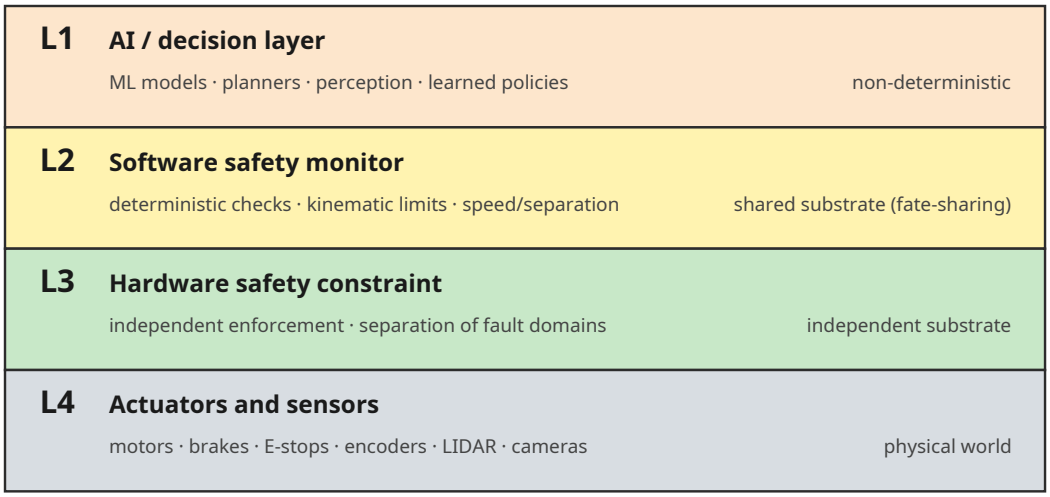


Figure 1. Physical AI Safety Stack: four horizontal layers labelled L1 AI / decision layer (ML models, planners, perception), L2 Software safety monitor (deterministic checks, kinematic limits), L3 Hardware safety constraint (independent enforcement, separation of fault domains), L4 Actuators and sensors (motors, brakes, E-stops, encoders).

The companion paper's conclusion — what it terms the **Software Safety Ceiling** — is that hardware-layer enforcement (L3) is needed when high assurance levels are required.

The Stack is not the only way to decompose a Physical AI system. Other taxonomies exist. The Stack has three properties that make it suitable for certification framing. First, it separates concerns by the type of failure each layer guards against — ML failure at L1, software failure at L2, hardware-attributable failure at L3, mechanical failure at L4. Second, it aligns with the existing standards landscape, which has long treated software safety (IEC 61508-3) and hardware safety (IEC 61508-2) as distinct concerns. Third, it permits per-layer certification with explicit treatment of cross-layer interaction.

2.2 Implications for Certification

Each layer of the Stack maps onto a different certification tradition:

Layer	Certification tradition	Standards
L1 — AI / decision	ML model evaluation; distributional-shift testing	ISO/IEC TR 5469 (emerging); ISO/IEC 22989; ISO/IEC 23053
L2 — Software monitor	Software safety integrity	IEC 61508-3
L3 — Hardware constraint	Hardware safety integrity	IEC 61508-2

L4 — Actuators / E-stop	Machinery safety	ISO 13849, ISO 13482, ISO 10218, ISO 12100

A certification framework for Physical AI must therefore address all four layers. It must also address their interaction — because failures in real systems cross layer boundaries. An ML model with a perception failure (L1) may pass through a software monitor that lacks the diagnostic to catch it (L2) and reach an actuator that has no hardware-level safe-state (L3). The interaction is the failure path.

2.3 Why Existing Standards Alone Are Insufficient

Existing standards address layers in isolation but lack four things needed for Physical AI:

1. **AI-specific evaluation criteria for L1.** No major standard yet contains a verified methodology for evaluating ML model safety integrity at SIL-equivalent rigor. Methods exist in academic literature; standards have not yet adopted them.
2. **Methodology for L1+L2+L3 interaction.** Standards address each layer separately. They do not yet address the interaction between an ML decision layer, a software monitor, and a hardware constraint, including the failure paths that cross layers.
3. **Common-cause failure analysis for ML+software hybrid channels.** IEC 61508-6 Annex D handles common-cause analysis for redundant hardware channels well. It does not handle the case where one channel is an ML model and another is software running on shared infrastructure — a case our forthcoming companion paper on common-cause failure analysis catalogues in detail.
4. **Distributional shift handling.** No standard yet defines what distributional shift detection a Physical AI system must implement, or what response protocol must follow detection. The concept entered standards discussion only with ISO/IEC TR 5469.

PAS-CF fills these four gaps. The remainder of this paper develops the framework.

---

3 Gap Analysis: Existing Standards

This section examines five major standards (or standard families). For each, we summarize what it covers and identify what it misses for Physical AI. The analysis is constructive: existing standards remain the foundation; PAS-CF augments rather than replaces them.

3.1 IEC 61508 — Functional Safety of E/E/PE Safety-Related Systems

Coverage. IEC 61508 is the seven-part baseline for functional safety. It covers random and systematic faults in electrical, electronic, and programmable-electronic

systems; classifies safety integrity into four levels (SIL 1 through SIL 4); specifies hardware safety integrity requirements (Part 2); specifies software safety integrity requirements (Part 3); and defines common-cause failure analysis methodology (Part 6, Annex D, including the β-factor model).

Gaps for Physical AI. Four gaps appear:

- No methodology for ML model safety integrity. The Part 3 software safety lifecycle assumes deterministic software behavior validated by static analysis, formal methods, or exhaustive test. ML models do not satisfy these assumptions.
- No criteria for distributional shift. The standard predates the concept.
- Software safety integrity assumes deterministic behavior. An ML model is a learned function whose output distribution may shift in deployment. Part 3 has no construct for this.
- The common-cause framework in Annex D is not directly applicable to ML-plus-software hybrid channels, where one channel is a learned function and another is hand-written software running on shared compute.

The practical consequence is that an ML-bearing safety function cannot today be assigned a SIL rating using IEC 61508 alone. Practitioners often work around this by claiming SIL on the L2 software monitor only and treating the ML layer as untrusted. This works to a point. It fails when the L2 monitor is not strong enough on its own — which is the Software Safety Ceiling case.

3.2 ISO 13849-1:2023 — Safety-Related Parts of Control Systems

Coverage. ISO 13849 specifies Performance Levels (PL a through PL e), categorical requirements for safety control circuits (Categories B, 1, 2, 3, 4), and diagnostic coverage requirements for safety-related control systems on machinery.

Gaps for Physical AI. Two gaps:

- AI-driven control loops are not contemplated. The standard assumes deterministic control logic.
- Diagnostic coverage assumes deterministic behavior. The notion of *coverage* requires that the diagnostic mechanism enumerate the failure modes it detects. ML failure modes resist enumeration.

ISO 13849 remains the working classification standard for safety-related control on European machinery. PAS-CF does not displace it; PAS-CF extends what evaluation it carries with AI-specific criteria.

3.3 ISO 13482:2014 — Personal Care Robots

Coverage. ISO 13482 addresses hazards specific to robots operating in human environments — mobile servant robots, physical assistant robots, person carrier robots. It defines risk-reduction measures for direct human contact.

Gaps for Physical AI. Three gaps:

- AI-driven hazards are not specifically addressed. The standard assumes the robot's behavior is the product of programmed rules.

- Adversarial input (deliberately constructed sensor input intended to cause misclassification) is not contemplated.
- Distributional shift in deployment environments is not addressed.

ISO 13482 is currently under revision. PAS-CF criteria may inform that revision; the integration question (where AI-specific criteria belong in ISO 13482's risk-reduction taxonomy) remains open.

3.4 ISO 10218-1:2025 — Industrial Robots

Coverage. The 2025 update to ISO 10218 addresses industrial robot safety, with substantial new provisions for collaborative operation (cobots) and updated requirements for safety-rated monitored stop, hand guiding, speed and separation monitoring, and power-and-force limiting.

Gaps for Physical AI. Two gaps:

- AI-driven decision-making receives partial coverage only. Where AI is referenced, the references are general.
- ML-specific failure modes are not addressed in detail.

The 2025 update closes some prior gaps. The remaining gap concerns vision-guided and ML-planned cobot operation, where the decision layer is an ML model.

3.5 ISO/IEC TR 5469 — Functional Safety and AI Systems

Coverage. ISO/IEC TR 5469 is a Technical Report (not yet a Standard) that addresses the combination of functional safety and AI systems. It identifies issues, sketches approaches, and defines terminology.

Gaps for Physical AI. Three gaps:

- It is a Technical Report, not a Standard. It carries informational status, not normative status.
- It addresses AI broadly. Physical AI — AI systems with actuators in the physical world — is not its primary scope.
- Implementation guidance is limited. The TR identifies where work is needed; it does not always say what the work should be.

ISO/IEC TR 5469 is the most directly applicable existing document to Physical AI Safety. PAS-CF is designed to be compatible with it and to provide implementation guidance the TR has not yet produced. The relationship is complementary: the TR identifies the questions; PAS-CF proposes evaluable answers.

3.6 The Synthesized Gap

Across the five families, existing standards address most of what Physical AI Safety requires. The remaining portion — the AI-specific concerns surfaced in Section 2.3 and the matrix below — requires new methodology. Figure 2 shows the structure of the gap.

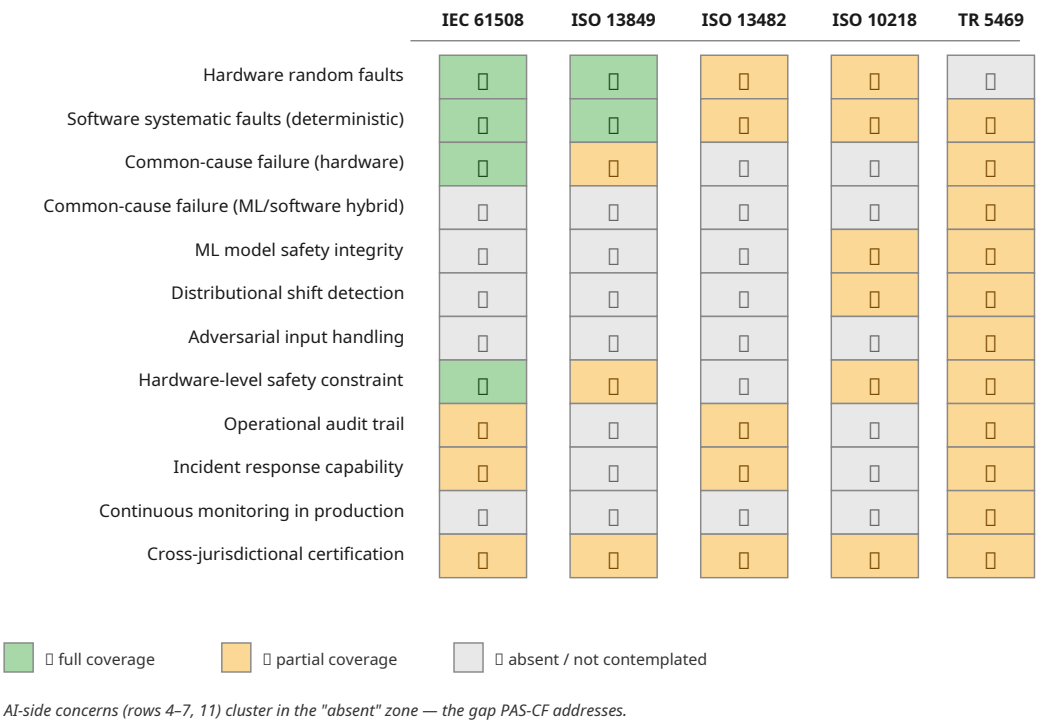


Figure 2. Standards coverage matrix: 12 rows of Physical AI Safety concerns (hardware random faults, software systematic faults, common-cause failure for hardware and for ML/software hybrid, ML model safety integrity, distributional shift detection, adversarial input handling, hardware-level safety constraint, operational audit trail, incident response capability, continuous monitoring in production, cross-jurisdictional certification) against 5 columns of standards (IEC 61508, ISO 13849, ISO 13482, ISO 10218, ISO/IEC TR 5469). Cell shading indicates full coverage (green), partial coverage (amber), or absent / not contemplated (light grey). AI-side concerns cluster in the absent band; hardware-side concerns are well covered.

Figure 2. Standards coverage matrix. Rows: Physical AI Safety concerns. Columns: existing standards. Cells indicate coverage: full (●), partial (◐), or absent (○).

The matrix surfaces the structure of the gap. Hardware-side concerns (random faults, hardware-level safety constraint) are well covered by IEC 61508 and ISO 13849. AI-side concerns — distributional shift, adversarial input, ML-software common-cause — appear only in ISO/IEC TR 5469, and there only partially. Operational concerns (audit trail, incident response) are covered partially across multiple standards but are not unified.

This is the gap PAS-CF addresses.

---

4 The Proposed Framework

4.1 Framework Overview

The Physical AI Safety Certification Framework, PAS-CF, is now defined.

> The **Physical AI Safety Certification Framework (PAS-CF)** is a proposed evaluation framework for Physical AI systems, integrating four AI-specific evaluation criteria with existing functional safety frameworks (IEC 61508, ISO 13849, ISO 13482). The framework provides a path for third-party certification of Physical AI systems against assurance levels comparable to those of legacy automation. PAS-CF is intended as a starting point for standards-body deliberation, not as a final standard.

PAS-CF augments existing functional-safety frameworks rather than replacing them. The augmentation takes the form of four evaluation criteria, each addressing one of the gaps identified in Section 3. Figure 3 shows the augmentation relationship.

Figure 3. PAS-CF as an augmentation layer. The four criteria sit above existing functional-safety frameworks rather than replacing them.

The four criteria are stated below, each followed by sub-criteria, rationale, and evidence expectations. The intent is that a third-party assessor — given the framework, the system specification, and the lifecycle artifacts — can return a profile per criterion (pass / partial / fail) with supporting findings.

4.2 Criterion 1 — AI Behavior Monitoring under Distributional Shift

> **Criterion 1 — AI behavior monitoring under distributional shift.** Evaluation of how the system detects and responds when operating conditions diverge from training/validation distribution. Requires: distributional shift detection mechanism, defined response protocol, evidence of testing across the operational design domain.

Sub-criteria.

- **1.1 Detection mechanism.** A documented method by which the system identifies that current input or operating conditions diverge from those represented

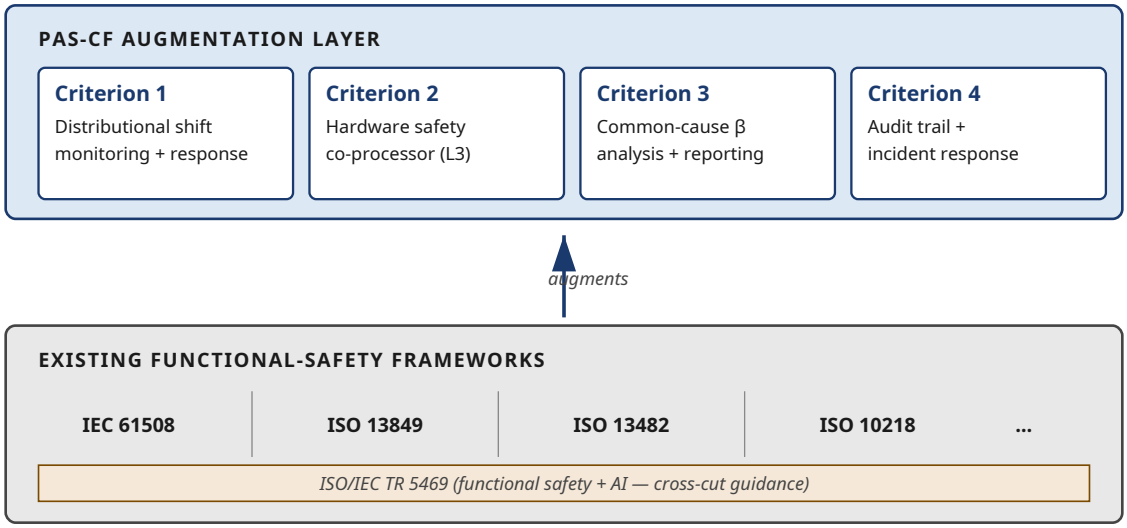


Figure 3. PAS-CF augmentation overview: a top blue band labelled PAS-CF Augmentation Layer contains four white criterion boxes (Criterion 1 distributional shift monitoring and response; Criterion 2 hardware safety co-processor; Criterion 3 common-cause beta analysis and reporting; Criterion 4 audit trail and incident response). An upward arrow labelled "augments" connects to a lower grey band labelled Existing Functional-Safety Frameworks containing IEC 61508, ISO 13849, ISO 13482, ISO 10218. A horizontal cross-cut bar at the bottom shows ISO/IEC TR 5469 functional safety plus AI cross-cut guidance.

in training and validation data. The method must be specified precisely enough that a third-party assessor can reproduce its operation.

- **1.2 Response protocol.** A defined response when shift is detected: graceful degradation to a safe behavioral envelope, fail-safe transition, or escalation to a human operator. The protocol must be deterministic and testable.
- **1.3 Operational design domain (ODD) coverage.** Evidence that the system was tested across the declared ODD, including deliberate excursions outside the ODD to confirm correct response.
- **1.4 Continuous production monitoring.** Distributional alignment is monitored during operation, not only during qualification. Drift indicators are logged.
- **1.5 Update process.** A documented process for retraining, revalidation, and recertification when distribution change exceeds a defined threshold.

Rationale. Distributional shift is the failure mode that distinguishes ML-bearing systems from deterministic ones. A traditional industrial controller behaves the same in deployment as in test, because its input space is bounded by physics and instrumentation. An ML-bearing system behaves correctly only as long as deployment inputs remain within the distribution it was trained on. Outside that distribution, its behavior is undefined. Criterion 1 turns this fact into an auditable expectation: the system must show how it detects when it has left the distribution it knows, and what it does then.

The operational design domain (ODD) is the contract. It states the conditions under which the system is qualified to operate — environment, inputs, tasks, edge cases. A well-specified ODD is precise enough that an assessor can construct a positive test (inputs inside the ODD, system behaves correctly) and a negative test (inputs outside the ODD, system enters its declared response state). Detection methods vary: statistical distance metrics on input distributions, confidence-thresholded classifier outputs, secondary out-of-distribution detectors trained alongside the primary model, or domain-specific physical plausibility checks. Criterion 1 does not mandate a single method. It requires that whichever method is used be documented, testable, and shown to behave correctly across the ODD boundary. The response protocol is equally important: silent failure is worse than detected failure that triggers a defined safe action.

Evidence base. A third-party assessor evaluating Criterion 1 receives: the ODD specification, the detector specification, test plans showing ODD coverage and out-of-ODD response, and operational logs showing the detector running in production. The assessor does not require deep ML expertise to confirm presence and operation of these elements. The criterion is verifiable.

4.3 Criterion 2 — Hardware Safety Co-Processor Presence and Integrity

> **Criterion 2 — Hardware safety co-processor presence and integrity.** Evaluation of Layer 3 (per the Physical AI Safety Stack from P1). Requires: documented hardware safety mechanism with separation of fault domains, verification of separation, evidence that Layer 3 enforces safety constraints independently of Layer 1 (AI) and Layer 2 (software monitor).

Sub-criteria.

- **2.1 Documented hardware safety mechanism.** A written specification of the L3 mechanism, its safety function, and the conditions under which it intervenes.
- **2.2 Separation of fault domains.** Evidence that the L3 mechanism is implemented with independence from L1 and L2: separate compute substrate, independent power supply, independent clock, separate memory, distinct supply chain where practicable.
- **2.3 Independence in enforcement.** Evidence that L3 enforcement actions (E-stop, motion limit, power cutoff) execute regardless of L1 or L2 state, including pathological L1/L2 states.
- **2.4 Independent supporting infrastructure.** Power, clock, and memory paths to L3 are documented and independent.
- **2.5 Architectural pattern.** The system identifies which architectural pattern it implements — in-path, side-monitor, or gateway, as described in our forthcoming companion paper on reference architecture — with rationale for the choice.

Rationale. Criterion 2 is the operational instantiation of the software safety ceiling argument from our forthcoming companion paper on that topic. A system that cannot show an L3 layer with separation of fault domains has no path to high assurance levels, because its software monitor shares fault modes with what it monitors. Hardware-layer enforcement, with documented separation, is the route past the ceiling.

Separation of fault domains is the central concept. Two channels share a fault domain if a single fault — a bug, a power glitch, a clock failure, a memory corruption, a supply-chain compromise — can disable both. Two channels are in separated domains if no single fault can disable both. Real separation is rarely absolute; it is graduated. An L3 mechanism running on dedicated silicon with its own power rail and its own clock has stronger separation than one running as a privileged process on the same processor as L1 and L2. The criterion asks the assessor to evaluate the strength of the separation against the assurance level claimed. Higher claims require stronger separation evidence. The criterion also addresses pathological L1 / L2 states: an L3 mechanism is only meaningful if it functions correctly when L1 and L2 are misbehaving — including deadlocked, looping, or producing high-rate but incorrect outputs.

Evidence base. The assessor evaluates: the L3 specification, block diagrams showing power, clock, memory, and bus separation, fault-injection test results targeting each separation claim, and evidence that L3 actions execute correctly under L1 and L2 fault conditions. Where redundancy is claimed, the evidence carries through to Criterion 3.

If the architectural-pattern reference work is not yet available at the time of assessment, sub-criterion 2.5 may be supported by equivalent industry-standard pattern descriptions, with the same set of expectations on documentation.

4.4 Criterion 3 — Common-Cause Failure Analysis with β Reporting

> **Criterion 3 — Common-cause failure analysis with β reporting.** Evaluation per IEC 61508-6 with explicit β-coefficient analysis (per P6) for any redundancy claims. Requires: documented common-cause analysis, β value with justification, hardware/software/team diversity evidence.

Sub-criteria.

- **3.1 Common-cause analysis per IEC 61508-6 Annex D.** A formal analysis identifying possible common-cause sources between redundant safety channels.
- **3.2 β-coefficient value with justification.** A reported β value with the analytical method behind it. The default of β = 0.1 (10 percent) used in some legacy analyses is not acceptable for ML-bearing channels without specific justification. A β-estimation methodology drawn from public information, suitable for this case, is the subject of a forthcoming companion paper.
- **3.3 Hardware diversity evidence.** Where redundancy is claimed, documentation of substrate, vendor, and supply-chain diversity between channels.
- **3.4 Software diversity evidence.** Where software redundancy is claimed, documentation of model architecture diversity, training data diversity, and inference path diversity.
- **3.5 Team diversity evidence.** Where redundancy is claimed at the development level, documentation that the redundant channels were specified, designed, and validated by independent teams.

Rationale. Common-cause failure analysis is the safety-engineering tool for evaluating redundancy claims. It asks: if Channel A fails, what is the probability that Channel B also fails for a related reason? In hardware practice, this analysis is mature. In ML-plus-software practice, it is largely absent. Two redundant ML models trained on the same dataset and running on the same compute substrate are not independent in the relevant sense — they fail together on inputs the dataset did not represent. Criterion 3 makes this analysis explicit and required where redundancy is claimed.

The β-coefficient default of 10 percent (β = 0.1), inherited from hardware practice, is the practical issue. In hardware, β = 0.1 corresponds to a meaningful claim: the two channels are mostly independent, with a 10 percent residual correlation captured by analysis of shared environment, supply chain, and design heritage. Applied to two ML models trained on the same dataset, β = 0.1 is unsupportable. The shared training distribution dominates correlated failure: both models will misbehave on the same out-of-distribution inputs. In such a case the effective β approaches 1.0 — the redundancy claim collapses. Criterion 3 requires the β value to be defended on its specific facts. Where defense fails, the redundancy claim is reduced or rejected, and the architecture is reconsidered.

Evidence base. The assessor reviews the common-cause analysis document, the β value and its derivation, and the documentation of hardware, software, and team diversity. The catalogue of common-cause sources in Physical AI from our forthcoming companion paper on common-cause failure analysis is the recommended starting point for the analysis. Where β cannot be defended at the value claimed, the redundancy claim is reduced or rejected.

4.5 Criterion 4 — Audit Trail and Incident Response Capability

> **Criterion 4 — Audit trail and incident response capability.** Evaluation of operational data capture, incident reporting protocol, and continuous monitoring. Requires: immutable audit log, documented incident response procedure, evidence of past incident handling (or simulated incidents).

Sub-criteria.

- **4.1 Immutable audit log of safety-relevant events.** Logging is tamper-evident, time-synchronized, and retained for a defined minimum duration. The events captured are specified.
- **4.2 Incident response procedure.** A written procedure covering detection, containment, root-cause analysis, corrective action, and external reporting.
- **4.3 Evidence of past incident handling.** Where the system has operated long enough to accumulate incidents, the incident record is part of the evidence base. Where it has not, simulated incidents (red-team exercises) substitute.
- **4.4 Lessons-learned integration.** Incidents and near-misses feed back into product development through a documented process.
- **4.5 External reporting.** Where regulators require reporting (machinery serious-incident reports under EU 2023/1230, AI Act incident reporting under EU 2024/1689), the reporting procedure is documented and tested.

Rationale. Criterion 4 turns operational reality into evidence. A system that runs in the field accumulates information about its own failure modes — every near-miss, every false-positive intervention, every handoff to a human operator. Without a capture mechanism, the information is lost. Criterion 4 requires the information be captured, analyzed, and acted upon. The criterion also recognizes that pre-deployment testing is bounded; post-deployment monitoring is what closes the loop.

The simulated-incident substitution addresses a practical reality: many systems applying for first certification have not yet operated long enough to accumulate real incidents. A red-team exercise — a planned scenario in which an incident path is exercised end-to-end through the response procedure — provides equivalent evidence for that case. The exercise must produce the same evidence trail as a real incident: detection, containment, root-cause analysis, corrective action, and external reporting where applicable. Audit-log retention is a separate concern: the criterion expects retention sufficient for forensic reconstruction of safety-relevant events months or years after the fact, with tamper-evidence to defeat the failure mode of after-the-fact log alteration. Time synchronization across distributed components matters when reconstructing a sequence of events that crosses subsystems. Criterion 4 treats these together because they fail together.

Evidence base. The assessor reviews: the audit-log schema and retention policy, sample log records showing safety-relevant events, the incident response procedure document, evidence of at least one incident or simulated incident handled end-to-end through the procedure, the lessons-learned integration record, and the external reporting procedure with evidence of test exercise.

4.6 Evaluation Output

The four criteria are applied in sequence. Figure 4 shows the resulting evaluation flow.

Figure 4. Four-criterion evaluation flow. The system is evaluated against each criterion in sequence. Output is a per-criterion assessment plus an overall PAS-CF profile.

The profile, not a single pass/fail verdict, is the output. A system may pass Criteria

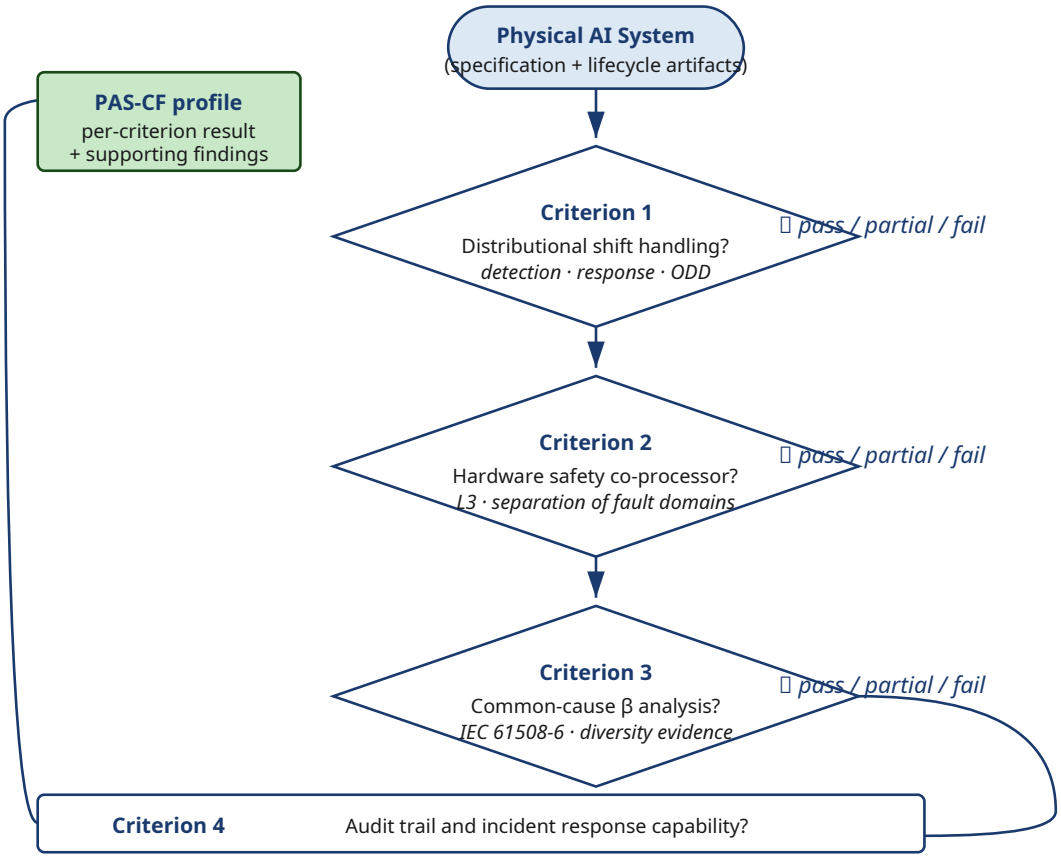


Figure 4. Evaluation flow: a Physical AI System input box at the top, with its specification and lifecycle artifacts, feeds downward into four diamond-shaped criterion gates evaluated in sequence — Criterion 1 distributional shift handling, Criterion 2 hardware safety co-processor presence, Criterion 3 common-cause beta analysis, Criterion 4 audit trail and incident response. Each gate produces a pass / partial / fail outcome. The outcomes feed into a PAS-CF profile box showing the per-criterion result plus supporting findings.

2, 3, and 4 while remaining partial on Criterion 1 — and that profile is informative to the user, the regulator, and the certifier. A profile-based output is also more honest about where Physical AI practice stands today: parts of the framework are easier to satisfy than others, and the profile makes that visible.

---

5 Mapping to Existing Standards

PAS-CF does not stand alone. Its criteria align with existing standards as follows.

> The framework maps to existing standards as follows: - **IEC 61508 SIL levels** → corresponding PAS Levels (per P4) - **ISO 13849 PL levels** → corresponding PAS Levels - **ISO 13482** (personal care robots) → application-specific PAS profile - **ISO 10218** (industrial robots) → application-specific PAS profile - **ISO/IEC TR 5469** (functional safety + AI) → directly applicable
>
> PAS-CF does not replace existing standards. It augments them with AI-specific criteria.

A PAS-Level maturity scale, modeled on the SIL/PL convention so that practitioners familiar with functional-safety classification can place a Physical AI system on a comparable scale, is introduced in a forthcoming companion paper on the maturity model. That companion paper provides a self-assessment route via questionnaire; PAS-CF provides the third-party-assessment route.

The criterion-by-standard mapping is given in Table 1.

Table 1. PAS-CF criterion mapping to existing standards.

PAS-CF criterion	IEC 61508	ISO 13849	ISO 13482	ISO 10218	ISO/IEC TR 5469
C1 — Distributional shift	Partial (system safety req.)	Not covered	Not covered	Partial	Yes (recently)
C2 — Hardware safety co-processor	Yes (61508-2)	Partial	Not covered	Partial	Partial
C3 — Common-cause β	Yes (61508-6 Annex D)	Partial	Not covered	Not covered	Partial
C4 — Audit trail / incident response	Partial (61508-1 lifecycle)	Not covered	Partial	Not covered	Partial

The implication of the mapping: PAS-CF can be implemented today as an extension to an IEC 61508 plus ISO 13849 baseline, with reference to ISO/IEC TR 5469 for AI-

specific guidance. No new infrastructure is required for adoption. The criteria are evaluable against existing functional-safety lifecycle artifacts, with additions specific to ML and L3 enforcement. Figure 5 places PAS-CF in relation to the surrounding standards.

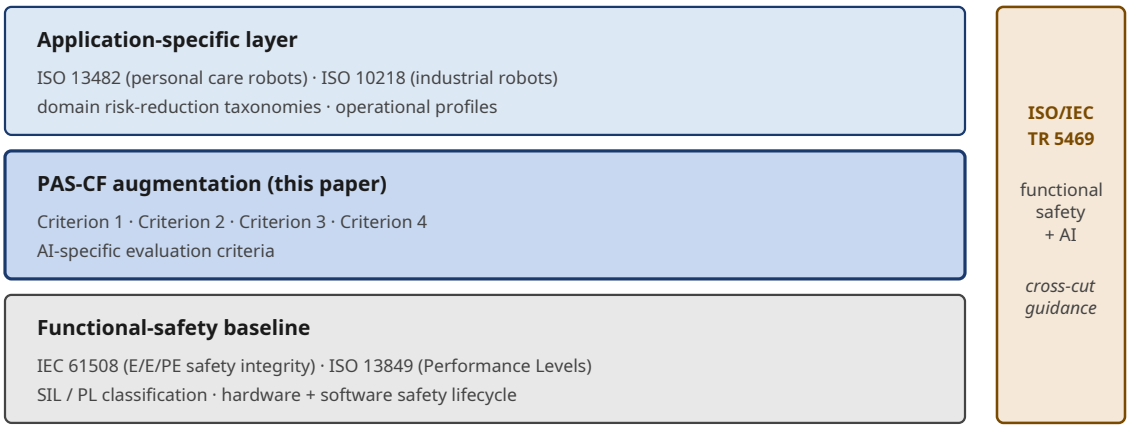


Figure 5. Standards integration view: three horizontal layers stacked vertically — the top blue layer Application-specific layer contains ISO 13482 (personal care robots) and ISO 10218 (industrial robots); the middle bold blue layer is the PAS-CF augmentation contributed by this paper containing Criteria 1 through 4; the bottom grey layer is the Functional-safety baseline containing IEC 61508 and ISO 13849. A vertical orange bar on the right runs the full height across all three layers and is labelled ISO/IEC TR 5469 functional safety plus AI cross-cut guidance.

Figure 5. Standards integration view. PAS-CF augments the IEC 61508 plus ISO 13849 baseline. Application-specific layers (ISO 13482, ISO 10218) determine which profile applies. ISO/IEC TR 5469 provides cross-cutting AI guidance.

The integration view answers an objection that may be anticipated: that PAS-CF duplicates existing work. It does not. PAS-CF is the augmentation layer. It contains what the existing structure does not yet contain — and only that.

---

6 Validation Cases

This section applies PAS-CF to three example Physical AI architectures. The cases use only public information or hypothetical configurations. They are illustrative, not certifying assessments. A real assessment would draw on internal documentation not available here.

6.1 Case 1 — Industrial Cobot with AI-Vision-Guided Picking

Configuration.

Layer	Implementation
L1	ML model for object detection and grasp planning
L2	Software safety monitor watching kinematic limits, joint torques, speed and separation
L3	Hardware E-stop and torque-limited drives, separate power and clock
L4	Cobot arm meeting ISO 10218-1 power-and-force limiting

PAS-CF profile (illustrative).

Criterion	Assessment	Notes
C1 — Distributional shift	Partial	Detection rudimentary; response protocol present but not validated across full ODD
C2 — Hardware safety	Pass	Independent E-stop with separation of fault domains
C3 — Common-cause β	Fail	β not analyzed; redundancy claims unsupported
C4 — Audit trail	Partial	Logging present; incident response not formalized

The case shows a typical industrial-cobot strength on the hardware side and weakness on the AI side. C2 is solid because machinery-safety standards already require it. C1 and C3 reveal the AI-side gap.

6.2 Case 2 — Autonomous Mobile Robot for Warehouse Logistics

Configuration.

Layer	Implementation
L1	ML for SLAM, navigation, and human-detection
L2	ROS2-based software safety nodes; speed and separation monitoring
L3	Variable across products — many AMRs in market lack a distinct L3
L4	Drive train, LIDAR, ultrasonics, bumpers

PAS-CF profile (illustrative).

Criterion	Assessment	Notes
C1 — Distributional shift	Partial	Many AMRs handle map-distribution shift; few handle perception-distribution shift
C2 — Hardware safety	Often weak	Where L3 is absent, the criterion fails
C3 — Common-cause β	Fail	β rarely analyzed in current AMR practice
C4 — Audit trail	Partial	Operational logs exist; incident response varies by operator

This case typically returns a maturity profile near the lower middle of the scale described in our forthcoming companion paper on the maturity model. C2 is the most common gap. Some AMR vendors are now retrofitting L3 mechanisms in response to certification pressure under EU 2023/1230.

6.3 Case 3 — Surgical Robot in Autonomous Mode

Configuration.

Layer	Implementation
L1	ML for tissue identification and trajectory planning
L2	Software safety with redundant validation channels
L3	Hardware safety constrained by FDA Class II/III process; force-limiting, watchdog timers
L4	Surgical end-effectors, encoders, force sensors

PAS-CF profile (illustrative).

Criterion	Assessment	Notes
C1 — Distributional shift	Partial	ODD tightly constrained by clinical indication; shift detection often patient-population-based

Criterion	Assessment	Notes
C2 — Hardware safety	Pass	FDA process drives hardware-side rigor
C3 — Common-cause β	Pass	β analyzed under FDA review process
C4 — Audit trail	Pass	Required by medical-device regulation

Surgical robots typically pass three of four criteria. The criterion they do not yet fully pass — C1, distributional shift — is the AI-specific gap that PAS-CF surfaces, and one that the medical-device regulatory regime is now beginning to address.

6.4 Cross-Case Patterns

Table 2 summarizes the patterns observed across the three example architectures.

Table 2. Cross-case patterns observed across the three example architectures.

Pattern	Frequency across cases
C2 (hardware safety) is the most common gap in the broader industry, though strong in regulated sectors	Two of three weaker; one strong
C3 (common-cause β) is rarely formally documented for ML-bearing channels	Three of three weak (only the regulated case partial-to-pass)
C1 (distributional shift) is the newest and least mature criterion	Three of three partial
C4 (audit trail) varies sharply by regulatory requirement	Three of three variable

The patterns are consistent with the empirical findings from our forthcoming companion paper on the Physical AI Safety Gap. The gap is not uniform. It is concentrated in C1 and C3, with C2 driven by sector regulation and C4 driven by external reporting obligations.

A final observation: regulatory pressure tends to close C2 and C4 first, because they map onto existing regulatory machinery. C1 and C3 require new methodology — which is why they belong in PAS-CF.

---

7 Implementation Roadmap

PAS-CF is intended for staged adoption. The roadmap below covers three horizons.

7.1 Short Term — 12 Months

The short-term horizon is voluntary adoption by industry and pilot use by certifying bodies.

Action	Actor	Outcome
Adopt PAS-CF as voluntary self-assessment instrument	Physical AI vendors	A baseline for internal safety review
Pilot PAS-CF as augmentation to existing certification	Notified bodies (TÜV SÜD, TÜV Rheinland, SGS, Pilz, others)	Practical experience with the criteria
Coordinate with the ISO/IEC TR 5469 working group	Researchers, framework authors, standards delegates	Alignment with the most directly relevant existing document
Publish a v2 of PAS-CF	Framework community	Refinement based on first-year lessons

7.2 Medium Term — 3 Years

The medium-term horizon is formal proposal to standards bodies.

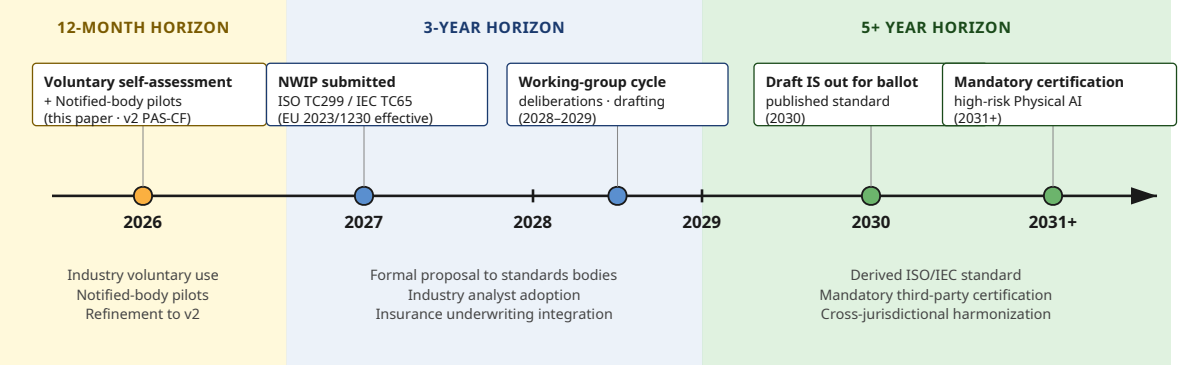
Action	Actor	Outcome
Submit a New Work Item Proposal (NWIP) to ISO TC299 (Robotics) and / or IEC TC65 (Industrial Process Measurement, Control and Automation)	National standards body delegates	Formal entry into the standards-development process
Incorporate PAS-CF criteria into harmonized standards under EU 2023/1230	EU notified-body process; harmonization committees	Presumption-of-conformity routes
Adoption of PAS-CF as a reference framework by industry analysts	Industry analyst firms	Market visibility of compliant suppliers
Adoption by insurance underwriters as a risk-pricing input	Insurance industry	Economic reinforcement of safety practice

7.3 Long Term — 5+ Years

The long-term horizon is a derived ISO/IEC standard with mandatory third-party certification for high-risk Physical AI.

Action	Actor	Outcome
Publish a formal ISO/IEC standard derived from PAS-CF	ISO TC299 / IEC TC65 working group	Normative framework
Mandatory third-party certification for high-risk Physical AI applications	Regulators (EU, US, Japan, Israel, others)	Enforceable assurance regime
Cross-jurisdictional harmonization (EU, US, Japan, Israel)	International coordination	Mutual recognition
Mature certification ecosystem	Notified bodies, training organizations, accredited assessors	Industry capacity

Figure 6 shows the implementation roadmap on a horizontal timeline.



The structure (voluntary formal mandatory) follows the established trajectory of IEC 61508 and ISO 26262.
Dates are illustrative; standards-body cycles vary.

Figure 6. Implementation roadmap timeline from 2026 through 2031+, divided into three coloured horizon bands — yellow 12-month horizon (2026: voluntary self-assessment plus notified-body pilots), blue 3-year horizon (2027 NWIP submitted to ISO TC299 / IEC TC65 alongside EU 2023/1230 effective; 2028–2029 working-group cycle), green 5+ year horizon (2030 draft International Standard out for ballot; 2031+ mandatory certification for high-risk Physical AI). Each year on the timeline carries a milestone marker. Below-axis annotations summarize the activities of each horizon: industry voluntary use, formal proposal to standards bodies, and derived ISO/IEC standard with mandatory third-party certification.

Figure 6. Implementation roadmap timeline. Voluntary self-assessment in 2026, formal proposal to standards bodies in 2027, working-group deliberation through 2028 and 2029, draft standard out for ballot in 2030, mandatory certification regimes from 2031.

The timeline is illustrative. Standards-body cycles vary; political and technical contingencies move dates. The structure — voluntary, then formal, then mandatory — is the established pattern for safety frameworks. The IEC 61508 trajectory followed approximately this shape over its initial decade. ISO 26262 (automotive functional

safety) followed a comparable path. The pattern is well-established; the contribution of this paper is to apply it to Physical AI early enough to be useful.

---

8 Limitations and Future Work

8.1 Limitations of PAS-CF v1.0

PAS-CF v1.0 is a starting point. We identify five limitations.

- **Built on currently-known failure modes.** The F1 through F6 failure modes catalogued in our forthcoming companion paper on the definitional framework reflect current understanding. New failure modes will emerge as Physical AI deployment broadens. The framework will require updating.
- **Distributional shift methodology still maturing.** Criterion 1 depends on a technical foundation that academic literature is still building. Detection methods, response protocols, and verification of detectors are active research areas. The criterion specifies what evidence is required without locking in a single methodology — a deliberate choice to permit evolution, but one that leaves implementation specifics to the assessor's judgment.
- **Hardware safety methodology assumes current architectural patterns.** Criterion 2 references three architectural patterns described in concurrent work — in-path, side-monitor, and gateway. New patterns may require extension of the criterion.
- **Cross-jurisdictional harmonization is partial.** The EU, the US, Japan, and Israel are converging on AI-bearing safety regulation, but they are not yet fully aligned. Mutual recognition between regimes is an open issue, and PAS-CF cannot resolve it unilaterally.
- **Voluntary adoption may be slow.** Without regulatory mandate, adoption depends on commercial incentive. Insurance and procurement levers are not yet fully engaged. The 12-month horizon in Section 7 may be optimistic if these levers do not engage.

8.2 Future Work

Five directions are identified for future work on PAS-CF and the surrounding ecosystem.

- **Empirical validation across more architectures.** The three cases in Section 6 are illustrative. A larger empirical study, applying PAS-CF to a representative sample of deployed systems, would strengthen the framework. This is candidate work for a forthcoming companion paper.
- **Extension to multi-agent and fleet certification.** PAS-CF v1.0 addresses single systems. Fleets of Physical AI systems present additional concerns — coordination failures, fleet-wide model updates, emergent behaviors at scale. Extension is needed.
- **Integration with cybersecurity frameworks.** Safety and security are no longer separable in connected Physical AI. Integration with IEC 62443, IEC 81001-5-1, and the EU Cyber Resilience Act is a needed extension.

- **Open-source reference implementations.** The field would benefit from open-source reference implementations of PAS-CF assessment instruments, audit-log schemas, and L3 architectural patterns. These would lower the barrier to adoption.
- **Integration with insurance underwriting frameworks.** PAS-CF profiles can serve as inputs to insurance pricing. Formalizing this linkage would create economic incentive for safety practice and would support the medium-term horizon in Section 7.

---

9 Open Research Questions

Beyond the limitations of PAS-CF v1.0, the field carries open research questions whose resolution will shape Physical AI Safety over the next decade. Table 3 lists ten such questions.

Table 3. Open research questions for the Physical AI Safety field.

#	Question	Relevance to PAS-CF
1	Quantitative ML safety integrity. What metric replaces SIL when the system contains ML components?	Underpins eventual SIL-equivalent classification of Physical AI
2	Common-cause failure analysis for ML+software hybrid channels. What is the methodology?	Strengthens Criterion 3
3	Verification of distributional shift detection. How is it confirmed that a detector catches shift?	Strengthens Criterion 1
4	Continuous-monitoring methodology. What is the operational data minimum required for assurance?	Strengthens Criteria 1 and 4
5	Cross-jurisdictional certification mutual recognition. What is the path?	Determines reach of any future PAS-CF-derived standard
6	Cost-effective Physical AI safety certification. Current cost may exceed startup capacity — what are the alternatives?	Determines accessibility of certification

#	Question	Relevance to PAS-CF
7	Open-source reference implementations. Should the field have these, and at what scope?	Determines pace of adoption
8	Insurance and certification integration. Where do the two meet?	Determines economic reinforcement
9	Multi-agent and fleet safety certification. What is the right unit of certification?	Determines extension scope
10	Continuous (versus point-in-time) certification. What is the model?	Reframes the certification act itself

These ten questions are the agenda for the next five years. The framework presented here is consistent with each of them; the answers will inform future versions of PAS-CF.

---

10 Conclusion

This paper has synthesized seven prior contributions on Physical AI Safety into a proposed certification framework. The synthesis was made possible by the fact that the prior contributions converge. Each is a piece of a coherent disciplinary structure. The definitional framework, the software safety ceiling argument, the architectural baseline, the maturity model, the empirical baseline, the common-cause failure methodology, and the regulatory mapping — each appears in its own forthcoming companion paper. Together they form the disciplinary ground on which PAS-CF stands.

PAS-CF is the proposed integration of that structure into a certification framework. It rests on four AI-specific evaluation criteria — distributional shift handling, hardware safety co-processor presence and integrity, common-cause failure analysis with β reporting, and audit trail with incident response capability. It maps onto existing standards rather than replacing them. It is implementable today as a voluntary instrument, proposable to standards bodies in three years, and capable of becoming a normative framework over a five-year horizon.

The audience for this paper includes ISO TC299 and IEC TC65 working group delegates, certifying bodies, regulators, industry, and researchers. We invite each:

- **Standards bodies.** Review the criteria; propose revisions; consider opening a New Work Item.
- **Certifying bodies.** Pilot the framework; report findings; help refine the criteria.
- **Regulators.** Consider PAS-CF as a candidate reference for harmonized standards under EU 2023/1230 and parallel regimes.

- **Industry.** Apply PAS-CF as voluntary self-assessment; share results.
- **Researchers.** Address the ten open questions of Section 9.

The Physical AI Safety Certification Framework is a starting point. The destination is a standard. The journey is the field's to travel together.

References

[1] International Electrotechnical Commission, "IEC 61508 (all parts) — Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems," 2010.

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

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

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

[5] International Organization for Standardization, "ISO 12100:2010 — Safety of Machinery — General Principles for Design — Risk Assessment and Risk Reduction," 2010.

[6] International Organization for Standardization and International Electrotechnical Commission, "ISO/IEC TR 5469 — Artificial Intelligence — Functional Safety and AI Systems."

[7] International Organization for Standardization and International Electrotechnical Commission, "ISO/IEC 22989 — Artificial Intelligence — Concepts and Terminology."

[8] International Organization for Standardization and International Electrotechnical Commission, "ISO/IEC 23053 — Framework for AI Systems Using Machine Learning."

[9] International Electrotechnical Commission, "IEC 62443 (series) — Industrial Communication Networks — IT Security for Networks and Systems."

[10] International Electrotechnical Commission, "IEC 81001-5-1 — Health Software and Health IT Systems Safety, Effectiveness and Security — Part 5-1: Security — Activities in the Product Life Cycle."

[11] European Union, "Regulation (EU) 2023/1230 of the European Parliament and of the Council of 14 June 2023 on Machinery (Machinery Regulation)," *Official Journal of the European Union*, 2023.

[12] European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)," *Official Journal of the European Union*, 2024.

[13] European Union, “Regulation (EU) 2024/2847 on Horizontal Cybersecurity Requirements for Products with Digital Elements (Cyber Resilience Act),” *Official Journal of the European Union*, 2024.

[14] National Institute of Standards and Technology, “Artificial Intelligence Risk Management Framework (AI RMF 1.0),” NIST AI 100-1, 2023.

[15] Anthropic, “Anthropic’s Responsible Scaling Policy,” 2023, with subsequent updates 2024.

[16] Stanford Institute for Human-Centered Artificial Intelligence (HAI), policy framework papers (various), Stanford University.

[17] Brookings Institution, AI policy papers (various).

[18] International Federation of Robotics, “World Robotics 2024–2025 Reports,” IFR.

[19] Published methodologies of TÜV SÜD, TÜV Rheinland, SGS, and Pilz on AI-bearing machinery safety assessment (various).

Mati Melchior · Independent Physical AI Safety Researcher mati@physical-ai-safety.com · physical-ai-safety.com 8-paper series on Physical AI Safety. P8 of 8.