Deployment intelligence for physical AI

Takeshi Ando, Panasonic R&D Center Singapore, Singapore, ando.takeshi@jp.panasonic.com

Abstract

Robotics is entering a foundation-model era, in which robot policies can generalize across tasks, objects and embodiments. Yet benchmark success does not show that a system is ready to act in factories, hospitals, warehouses or public spaces. Deployment requires a different kind of evidence: whether embodied intelligence remains valid, safe and accountable after its assumptions, sites and models change. This Perspective proposes deployment intelligence as a measurement science for that evidence. Rather than replacing task success, deployment intelligence extends evaluation to the model-in-system: the coupled policy, body, site, human workflow, safety infrastructure and governance process. We distinguish capability evidence from deployment evidence, define a readiness vector for physical AI, extend operational design domains into dynamic site envelopes and propose deployment-evidence cards as a reporting norm. The goal is to make deployability measurable before robots are treated as real-world infrastructure.

Keywords: physical AI; robot foundation models; vision-language-action models; deployment readiness; measurement science; safety assurance; robot learning; embodied AI

The missing layer between benchmark success and real-world readiness

Robotics has begun to absorb the foundation-model logic that transformed language and vision. Robot Transformers, vision-language-action (VLA) models and robot foundation models (RFMs) show that broad data, language-conditioned policies and cross-embodiment datasets can produce policies that generalize across tasks, objects and platforms.[1-8] Existing robot-learning benchmarks, from Meta-World, CALVIN, ManiSkill, LIBERO and BEHAVIOR-1K to DROID, BridgeData V2, Open X-Embodiment and ASIMOV, largely ask whether a policy succeeds, generalizes or refuses unsafe instructions under specified evaluation conditions.[9-16] They do not systematically ask whether the same system can be commissioned, transferred across sites, updated without regression, monitored at its cognitive boundary or governed after incidents.

The point is not that current benchmarks are flawed. They have clarified competence. The gap is that competence is only one type of evidence. A field robot must also provide deployment evidence: evidence that it can be commissioned, recovered, updated, secured, governed and maintained under site-specific constraints. Without this layer, the community risks mistaking benchmark progress for deployability.

The scientific question should therefore shift from “Can the robot do the task?” to “What evidence is sufficient to let embodied AI keep acting in the world?” This Perspective proposes deployment intelligence (DI) as the missing measurement layer for physical AI. DI is not a replacement for task success, nor a generic operations checklist. It is the boundary-aware, evidence-preserving intelligence of the model-in-system: the capacity to recognize, preserve and update the conditions under which embodied actions remain valid. This shift is timely because generalist robot policies are beginning to blur the boundary between model evaluation and system evaluation. As policies become more semantic, updateable and cross-embodied, the old separation between “model performance” and “deployment conditions” becomes untenable. A physical AI policy is not merely a predictor deployed in the world; it is an action-generating component whose validity depends on bodies, sites, people, updates and evidence trails.

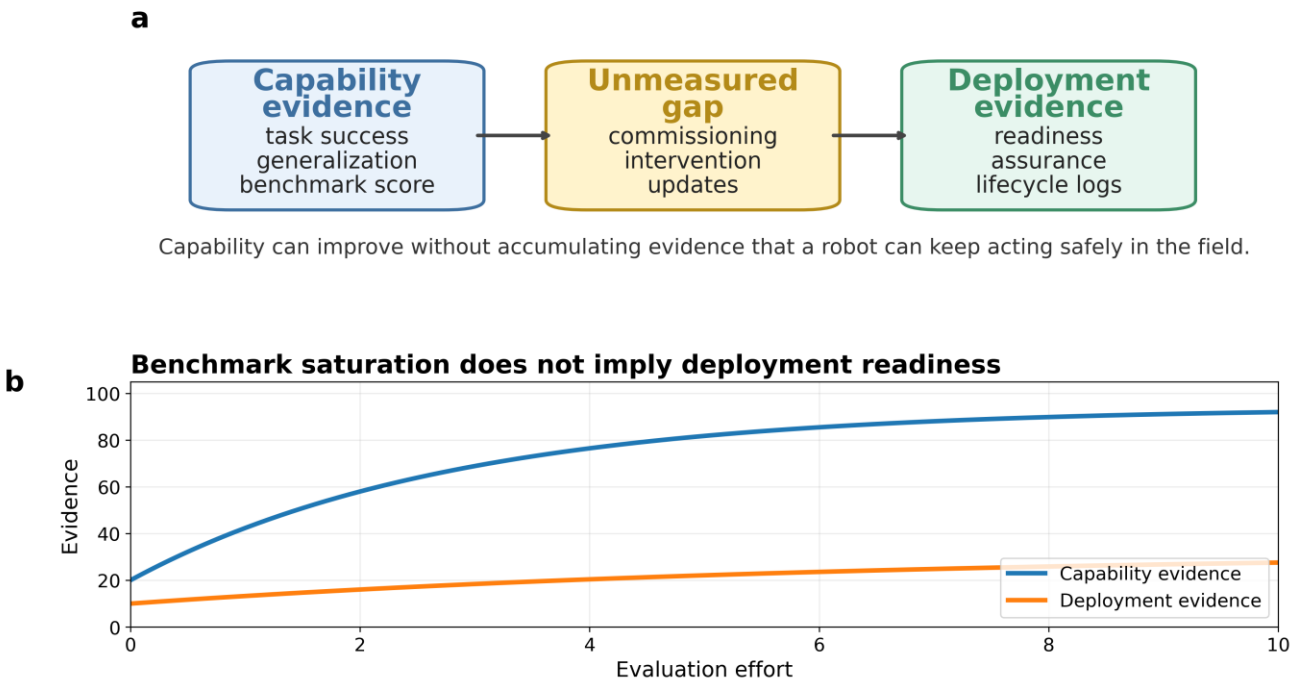

Fig. 1 | Capability evidence is not deployment evidence. a, Existing benchmarks primarily generate evidence about task competence, generalization and semantic safety. Real-world operation additionally requires evidence about commissioning, intervention, updates, assurance and lifecycle logs. b, Capability scores may improve and saturate while deployment evidence remains weak, creating a gap between benchmark success and field readiness. The curves are conceptual, not empirical trends.

DI also differs from existing readiness and reporting frameworks. Technology Readiness Levels track maturity; model cards and datasheets report model and dataset properties; AI risk-management frameworks and assurance cases organize risk and evidence; ODD formalizes operating conditions.[25-33] DI is needed because generative, updateable and embodied models act in an irreversible physical world. It connects model behaviour, robot embodiment, runtime assurance, site conditions, human intervention, update traceability and accountability evidence.

This comparison is important, because DI should not appear as a new name for deployment management. Its novelty is the coupling it makes explicit: model behaviour, physical embodiment, site assumptions and lifecycle evidence become one measurement object. That coupling is specific to physical AI. A text-only model can be evaluated without asking whether a manipulator damaged an object, whether a fleet update invalidated an old safety case, or whether a hospital workflow changed the meaning of a task request. A physical AI system cannot.

Prior frameworks were largely designed for systems whose behaviours were bounded by explicit specifications, fixed operating assumptions or static reporting artefacts. What is new about VLA/RFM-controlled physical AI is not complexity alone, but the combination of semantic generativity, continuous updateability and physical irreversibility in a single system. A framework that does not jointly measure these properties can remain useful within its original scope, yet still be structurally incomplete for evaluating deployability.

Table 1 | Why existing frameworks are necessary but insufficient for physical AI deployment. The comparison clarifies that DI is not a relabelling of TRL, ODD, model cards or safety cases, but a measurement layer that connects model behaviour, embodiment, site conditions, lifecycle updates and accountability evidence.

Existing framework	What it captures	What it misses for VLA/RFM-controlled physical AI	What DI adds
Technology Readiness Levels	Maturity of a technology from concept to operational use	Boundary-aware action validity, human intervention burden, model-update regression and site-specific evidence	A readiness vector grounded in deployment evidence, not only technology maturity
Model cards and datasheets	Model or dataset properties, intended use, limitations and evaluation conditions	Embodiment, facility integration, runtime recovery, ODD/site-envelope metadata and physical consequences of action	A reporting standard for embodied model-in-system behaviour
AI risk-management frameworks	Risk identification, documentation, governance and organizational controls	Concrete robotics telemetry: near misses, interventions, update rollbacks, safety stops and boundary violations	Operational evidence connecting AI governance to physical action

Safety cases and runtime assurance	Structured safety arguments, hazards, monitors and fallback mechanisms	Generative semantic action proposals, prompt-induced behaviour changes and post-update safety regression	Non-compensatory gates and lifecycle evidence for foundation-model behaviour
Operational Design Domain	External operating conditions for an automated system	Workflow/API boundaries and cognitive limits of a VLA/RFM or skill library	A site envelope that includes spatial, operational and cognitive boundaries

Why physical AI changes what evaluation means

Digital AI evaluation can often separate the model from its operating environment. A language-model benchmark may measure knowledge, reasoning or safety refusal without asking whether an actuator overheats, a map drifts, a worker becomes overloaded or a safety case becomes stale. Machine-learning evaluation and model documentation have developed important tools for reporting model behaviour, datasets and risk, but physical AI forces these tools to be coupled to bodies, sites, operators and lifecycle evidence.[26-33,41]

This makes deployment failure qualitatively different from ordinary performance degradation. In field robotics, long-term deployments repeatedly show that autonomy is constrained by persistence, recovery, charging, localization, network quality, intervention and maintenance, not by isolated task competence alone.[17-24] Human-automation and socio-technical research further shows that automation often shifts work from continuous control to rare, cognitively demanding interventions, creating hidden labour and fragile responsibility allocation.[46-53]

Three field patterns make the point concrete. A hospital logistics robot may navigate a corridor but fail deployment because elevator access, staff routines, patient privacy, congestion and information-system interfaces determine whether deliveries can scale.[20] A warehouse AMR may show robust navigation but become uneconomic if traffic deadlocks, charging windows, integration and human exception handling dominate throughput.[24] A public-space service robot may work technically but fail socially if local managers and users cannot interpret, repair or trust its behaviour over time.[18,21]

VLA/RFM-controlled robots amplify these issues. Their value comes from generality, semantic interpretation and updateability. The same properties create new deployment risks. A prompt, perception confidence, object label, tool library or model update may change not only task success, but also action timing, false stops, human workload, energy use, semantic safety and liability.[16,40] Existing robot standards and safety-control methods provide foundations for mechanical, collaborative and functional safety, but foundation-model robots also require semantic and operational safety: whether a high-level instruction is valid in the current site envelope and whether update evidence remains valid after a model, prompt, map or skill-library change.[34-40]

Deployment intelligence as a readiness vector

We define deployment intelligence as the measurable capability of a physical AI model-in-system to recognize, preserve and update the conditions under which its embodied actions remain valid, safe and accountable. Operationally, this includes whether the system can be commissioned, integrated, operated, recovered, updated, secured, governed and economically and organizationally sustained within a specified operational boundary. The unit of evaluation is not the model alone but the coupled stack of policy, robot body, site, human workflow, safety infrastructure and governance process.

DI should be reported as a readiness vector rather than a single universal score. Four layers are useful. The core capability layer retains task competence. The technical and resource layer measures commissioning, reliability, edge constraints and site-envelope behaviour. The socio-organizational layer measures hidden human labour, integration, economics and organizational adoption. The assurance and trust layer measures safety, security, governance and update readiness.

Fig. 2 | A layered readiness vector for deployment intelligence. DI is organized into four layers. Task competence remains necessary, but technical-resource, socio-organizational and assurance-trust dimensions are needed to evaluate whether capability can become sustained deployment. Critical dimensions such as safety, site-envelope, security, governance and updates should be treated as non-compensatory gates.

The reason for this structure is practical and scientific. A robot can be competent but not maintainable, safe but not economically deployable, integrated but update-fragile, or commercially attractive but impossible to govern after incidents. A single aggregate score would conceal these failure modes. A readiness vector exposes them.

Some dimensions are inevitably correlated. Reliability and human intervention, for example, are linked; safety and site-envelope behaviour are also linked. DI should therefore be treated as a hierarchical construct rather than as a set of independent variables. The lower-level metrics are observations; the higher-level layers are latent readiness factors. This framing permits measurement validation by expert review, inter-rater reliability, factor analysis and predictive validity testing, rather than relying on an arbitrary list of indicators.[42-45]

Table 2 | Deployment intelligence dimensions and examples of evidence. The metrics are examples rather than a universal checklist; domain-specific weighting and gate thresholds are required. Human-intervention evidence should include not only frequency, but also recovery complexity and operator cognitive workload.

Layer	Dimensions	Examples of deployment evidence
Core capability	Task competence	Task success, completion time, long-horizon completion and intervention-free success.
Technical and resource	Commissioning; operational reliability; edge/resource; site envelope	Setup time, calibration attempts, mean time between failures, recovery time, P99 latency, energy per successful task, thermal throttling, boundary detection latency and safe degradation.
Socio-organizational	Human intervention; integration; commercial; organizational	Human-intervention rate, teleoperation duration, recovery context cost, custom-interface count, workflow-integration effort, cost per task, support response time, training hours and role clarity.
Assurance and trust	Safety; security; governance; update	Near misses, unsafe-instruction rejection, secure update verification, incident-reporting readiness, traceability completeness, regression-test coverage, rollback success and assurance-case freshness.

The socio-organizational layer deserves emphasis because it is often invisible in robot-learning papers. A system can report few collisions while depending on a trained engineer who restarts it repeatedly, an operator who rewrites prompts after semantic failures, or a facilities manager who negotiates access before each run. Such labour is real, scales with deployment and may not transfer to a new site. Without measuring it, autonomy can shift responsibility onto humans as a moral crumple zone rather than reducing operational burden.

A compact formulation makes the non-compensatory logic explicit without turning this Perspective into a scoring manual. Let the normalized evidence score of DI dimension j be s_j, and let $R_{critical}$ denote safety, security, governance, update readiness and site-envelope awareness. A deployability judgement should not be a simple average. Deployment readiness can instead

be represented as a gated score Ψ, in which critical failures cap or block readiness regardless of high task competence elsewhere, as shown in equation (1).[30,34-41]

$$\Psi = \left[\prod_{g \in R_{critical}} \mathbf{1}(s_g \geq \theta_g)\right]\left[\sum_{j \notin R_{critical}} w_j s_j\right] \quad (1)$$

Here, θ_g is the domain-specific minimum evidence threshold for critical gate g, w_j is a domain-specific weight for non-critical dimension j, and $I(\cdot)$ is an indicator function. The equation is not a universal numerical threshold; it makes explicit a systems-safety principle. A robot can be competent and still not deployable if it lacks safety evidence, cyber-physical security, update traceability, governance or site-envelope awareness.

The critical set is selected by a simple principle: a dimension becomes a non-compensatory gate when its failure can invalidate deployment regardless of task competence because it creates unacceptable physical harm, uncontrolled operation outside the intended boundary, cyber-physical compromise, unapproved model change, legal non-compliance or loss of accountability. Commercial viability and maintenance burden may be severe, but they usually degrade deployment value rather than immediately invalidating the legitimacy of deployment.

From ODD to dynamic site envelopes

A central concept for deployment is the Operational Design Domain (ODD), widely used in automated driving to define where and when a system is intended to operate.[31,32] Physical AI needs a broader construct: the site envelope. A site envelope includes conventional external conditions such as space, lighting, weather and obstacles, but also workflow rules, infrastructure APIs, network quality, human density, data ownership, semantic permissions, skill-library preconditions and model-cognitive boundaries.

This matters because many physical AI failures occur at the boundary between the physical, operational and cognitive worlds. A hospital delivery robot may be geometrically inside its map but outside its site envelope when elevators are overloaded, staff workflows have changed or privacy constraints prevent data collection.[20] A manipulator may be physically able to grasp an object but outside its model-cognitive envelope if the VLA cannot distinguish a prohibited item from an ordinary one.[16] A warehouse AMR may remain collision-free but fail deployment if an API timeout prevents workflow completion.[24]

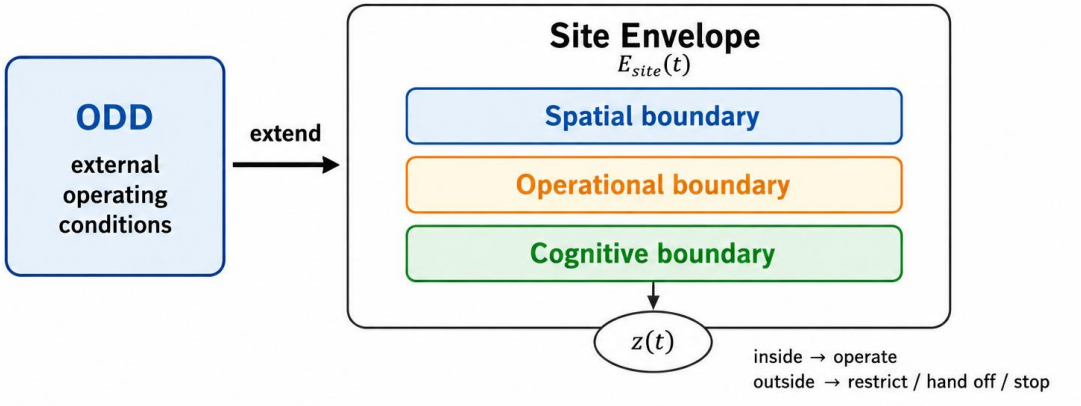

Fig. 3 | From Operational Design Domain to site envelope. Physical AI requires boundary monitoring across spatial, operational and model-cognitive states, not only across external environmental conditions. Readiness means detecting boundary violations and degrading safely before physical harm occurs.

Formally, the site envelope at time t can be written as a product-state constraint over three dynamic boundary sets, as shown in equation (2). The cognitive term is the physical AI-specific addition: it captures context-window limits, semantic uncertainty over objects or instructions, skill-library preconditions, payload and actuator limits, permissions and prohibited actions.

$$E_{site}(t) = \Omega_{spatial}(t) \times \Omega_{operational}(t) \times \Omega_{cognitive}(t) \quad (2)$$

When the system state $z(t)$ falls outside $E_{site}(t)$, a deployment-ready system should detect the violation and move toward restricted autonomy, human handoff or a minimum-risk condition before physical harm occurs.[16,37,38,54,55] Site-envelope readiness

should therefore measure not only whether a robot stays within a defined domain, but how quickly it detects boundary violations and how safely it degrades.

The most important data flywheel for deployable robotics may not be only successful trajectories. It may be the systematic logging of failures, interventions, calibration drift, maintenance actions, near misses, update regressions and workflow friction. Success data teaches what a robot can do; deployment data teaches what prevents it from remaining useful.[17-24,41]

From benchmarks to lifecycle evidence

How should the community measure DI without asking every laboratory to run multi-year field trials? We propose DI-Bench as a lightweight layer on top of existing benchmarks. DI-Bench does not replace Meta-World, LIBERO, DROID, Open X-Embodiment or ASIMOV.[9-16] It adds lifecycle phases around them: setup, relevant environment, single-site pilot, site transfer, model update, degraded condition and continuous audit.

Each phase collects different evidence. Setup measures how much effort is required before the first safe task. A relevant-environment phase measures performance under realistic but bounded conditions. A single-site pilot measures repeated operation, human intervention and reliability. Site transfer measures whether deployment scales beyond one customized location. Update regression tests whether a new model, prompt, skill library, map or safety monitor preserves competence, timing and safety. Degraded-condition testing probes network loss, congestion, blocked routes, sensor drift, low battery, staff shortage or unexpected human behaviour.[17-24,56-59]

Readiness levels can be useful as an optional interpretation of DI-Bench evidence, but they are not a second framework: DI is the construct, DI-Bench is the evidence protocol and readiness levels are merely a communication layer. Supplementary Information provides one possible mapping between evidence phases and physical AI readiness levels.

Making readiness testable rather than rhetorical

A common weakness of readiness frameworks is that they become lists of desirable properties. DI should avoid this by being testable. Three validation questions are central. First, content validity: do the dimensions cover the main sources of deployment friction? Second, construct validity: do the collected logs and indicators actually measure the intended readiness dimensions? Third, predictive validity: do early DI profiles predict future outcomes such as human intervention, downtime, incidents, rollback, operator acceptance or discontinuation?[42-45]

Predictive validation can be simple. Suppose a DI profile is collected at the start of a pilot. Future human interventions can be modelled as count data, with operational exposure as an offset; time to serious failure or withdrawal can be modelled with survival analysis.[65,66] The claim is falsifiable: if systems with higher integration, update or site-envelope readiness do not show fewer future interventions or failures, then the dimensions, weights or log schema must change.

Construct validity also depends on avoiding garbage-in, garbage-out. Logs should not rely only on self-reported post-hoc labels. Wherever possible, DI evidence should be anchored in objective telemetry: time-stamped robot state, controller state, middleware lifecycle events, diagnostics, audit trails, secure update records and standardized incident schemas.[56,57] Subjective measures such as workload, trust and workflow disruption remain important, but they should be paired with validated instruments and inter-rater reliability rather than treated as informal impressions.[42-49]

A reporting standard for physical AI

The most immediate intervention we propose is not a new benchmark, but a change in reporting norms. This is deliberately a low-friction proposal. Even when a study cannot measure long-term autonomy, site transfer or update regression, it can still report that those evidence classes were not evaluated. Negative reporting changes nothing about an algorithmic contribution; it changes what readers are allowed to infer from it. An unmeasured dimension is not a failed dimension, but an honest boundary on a deployment claim. Today, many robot-learning papers report task success, generalization splits and qualitative videos.[9-16] A DI-aware report would keep these measures but add a compact deployment-evidence card, analogous in spirit to model cards and datasheets but extended to embodied, site-dependent systems.[26,27] Such a card would state the robot embodiment, site or simulated site, duration of evaluation, number of trials, number and type of human interventions, setup and calibration effort, compute platform, P95 and P99 latency, energy or power envelope, update status, safety monitor used, fallback mode and known exclusions from the site envelope.

Table 3 | Minimal deployment-evidence card for physical AI papers. This template turns DI into a reporting standard analogous in spirit to model cards and datasheets, while preserving partial compliance and negative reporting.

DI layer	Mandatory reporting protocol	Reported state / evidence path
Core capability	Intervention-free success rate; completion time	Report measured value or state “not

	P50/P95; long-horizon step depth.	measured".
Technical and resource	Setup time; calibration attempts; P99 latency; energy per task; thermal-throttling onset; offline survival time.	Report platform, site assumptions, objective telemetry and evidence path.
Socio-organizational	Human-intervention rate; teleoperation time; recovery context cost; operator workload; custom API count; workflow-integration effort; training hours.	Report operator role, hidden labour, cognitive recovery burden and validated workload instrument where available.
Assurance and trust	Unsafe-instruction rejection; boundary-detection latency; external runtime-assurance monitor; regression-test coverage; rollback success; audit-log completeness.	Report passed gates, failed gates, excluded gates and whether evidence came from telemetry, tests or expert judgement.
Negative reporting	Explicitly declare dimensions unmeasured, unmodelled or outside the study.	Example: long-term cyber-security and multi-site organizational adoption were not evaluated.

Table 3 is intended as the practical reporting interface for this Perspective: it asks authors to report measured evidence where available and to state explicitly when evidence is not measured. This would not require every paper to run production pilots. It would require authors to distinguish what has been demonstrated from what remains unmeasured. Negative reporting prevents readers from inferring deployment readiness from the absence of evidence.

This reporting change is deliberately modest. It does not require a new community-wide benchmark before improving the literature. It asks authors to make the evidential boundary of a paper visible. If a policy is evaluated only in simulation, the card makes that clear. If a real-robot study does not measure energy, update regression or human intervention, the absence becomes visible rather than being silently absorbed into claims of generality. In this way, DI can improve transparency immediately while longer-term DI-Bench protocols mature.

The same principle applies to models marketed as generalist. Generality should be separated into at least three claims: semantic generality, embodiment generality and deployment generality. A system may be strong on the first and weak on the third. Collapsing these claims into a single "generalist robot" narrative obscures the evidence needed by researchers, users, regulators and insurers.[1-8]

Near-term research agenda

Over the next phase of foundation-model robotics, the field should aim for physical AI systems whose competence is accompanied by deployment evidence. Model builders should report not only task success and generalization but also latency, energy, semantic refusal, update regression and failure modes. Dataset builders should log not only demonstrations but also interventions, near misses, calibration, site conditions and recovery. Benchmark designers should include setup, site transfer and degraded-condition phases. Safety researchers should connect semantic safety with runtime assurance and dynamic safety cases. Standards bodies should begin specifying logs, model-update evidence and site-envelope documentation for AI-generated robot behaviour.[16,28-40,68]

This agenda should be viewed as a research programme rather than a compliance demand. Different papers can contribute different evidence classes. Algorithmic papers may begin with negative reporting and instrumented latency or energy measurements. Dataset papers can add intervention and failure metadata. Field studies can contribute site-transfer and recovery evidence. Standards bodies and regulators can later use the same vocabulary to specify which evidence classes are required for high-risk deployment claims.

To avoid making DI an entry barrier that only large organizations can satisfy, the community should build low-cost virtual DI toolchains. However, low-cost does not mean cost-free. Virtual DI toolchains require simulator instrumentation, latency and packet-loss models, thermal and energy telemetry, site-envelope metadata, versioned update tests and partial calibration against field data. They should therefore be treated as pre-validation tools that make deployment assumptions explicit, not as substitutes for field evidence.[56-59]

The most valuable deployment data may be rare, negative and operational: why the robot stopped, who intervened, what changed after an update, which site rule was violated, how long recovery took and whether trust was repaired.[17-24,46-53] These data are often personal, commercially sensitive or security-relevant. A practical DI infrastructure should therefore support privacy-preserving cross-site validation through techniques such as federated learning, differential privacy and secure aggregation.[60-64]

Limitations and claim discipline

This Perspective makes a narrow claim. It does not argue that deployment readiness is more important than model capability, nor that every laboratory study must become an industrial field trial. It argues that the evidence required for deployability is

different from the evidence produced by current capability benchmarks, and that this difference should be made explicit, measurable and testable.

The framework has limitations. DI does not yet have large-scale predictive validation. It does not define universal thresholds across domains. Some dimensions, especially human intervention, trust repair, organizational readiness, workflow burden and recovery context cost, require validated measurement instruments and inter-rater reliability. Cognitive-boundary monitoring is difficult because VLA policies may be overconfident outside their true competence. Field logs may contain personal data, customer processes, facility maps and commercially sensitive failure information. These limitations do not weaken the need for DI; they define the research programme required to make DI scientifically credible and equitable.

Conclusion

Robotics foundation models are changing what robots can attempt. Deployment intelligence asks what evidence is sufficient to let embodied AI keep acting in the world. This distinction will become more important as robots move from demonstrations to factories, hospitals, homes, farms, infrastructure and public spaces.

The field should therefore expand evaluation from benchmark success to lifecycle evidence. A robot policy should be judged not only by whether it completes a task, but by whether it can be installed without heroic effort, operate without hidden human labour, detect when it is outside its site envelope, degrade safely, update without regression, leave auditable evidence and remain economically and organizationally sustainable.

The next generation of physical AI will require more than larger models. It will require a measurement science for deployability: a way to make boundary awareness, evidence preservation, safe degradation and negative reporting as normal as task success. Deployment intelligence offers a candidate language for that science.

Data availability

No new empirical dataset was generated for this Perspective. The article synthesizes information from published literature and proposes a conceptual measurement framework. Future DI-Bench validation should release anonymized deployment logs, benchmark metadata and scoring code whenever legal, privacy and commercial constraints permit.

Code availability

No executable code is required to reproduce the arguments in this Perspective. Future DI-Bench implementations should release machine-readable log schemas and scoring scripts where possible.

Author contributions

T.A. conceived the perspective, developed the central argument and drafted the manuscript.

Competing interests

T.A. declares no competing interests.

Funding

T.A. declares no specific funding for this work.

References

1. Brohan, A. et al. RT-1: Robotics Transformer for real-world control at scale. Preprint at https://arxiv.org/abs/2212.06817 (2022).
2. Brohan, A. et al. RT-2: Vision-Language-Action models transfer web knowledge to robotic control. In Conference on Robot Learning (2023).
3. Open X-Embodiment Collaboration. Open X-Embodiment: robotic learning datasets and RT-X models. In Proc. IEEE Int. Conf. Robot. Autom. (2024).
4. Kim, M. J. et al. OpenVLA: an open-source Vision-Language-Action model. Preprint at https://arxiv.org/abs/2406.09246 (2024).
5. Octo Model Team. Octo: an open-source generalist robot policy. Preprint at https://arxiv.org/abs/2405.12213 (2024).
6. Black, K. et al. π0: a vision-language-action flow model for general robot control. Preprint at https://arxiv.org/abs/2410.24164 (2024).

7. Black, K. et al. π0.5: a Vision-Language-Action model with open-world generalization. Preprint at https://arxiv.org/abs/2504.16054 (2025).
8. Bjorck, J. et al. GR00T N1: an open foundation model for generalist humanoid robots. Preprint at https://arxiv.org/abs/2503.14734 (2025).
9. Yu, T. et al. Meta-World: a benchmark and evaluation for multi-task and meta reinforcement learning. In Conference on Robot Learning (2020).
10. Mees, O. et al. CALVIN: a benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. IEEE Robot. Autom. Lett. 7, 7327-7334 (2022).
11. Mu, Y. et al. ManiSkill2: a unified benchmark for generalizable manipulation skills. In International Conference on Learning Representations (2023).
12. Liu, B. et al. LIBERO: benchmarking knowledge transfer for lifelong robot learning. In NeurIPS Datasets and Benchmarks Track (2023).
13. Srivastava, S. et al. BEHAVIOR-1K: a human-centered, embodied AI benchmark with 1,000 everyday activities and realistic simulation. In Conference on Robot Learning (2022).
14. Khazatsky, A. et al. DROID: a large-scale in-the-wild robot manipulation dataset. In Robotics: Science and Systems (2024).
15. Walke, H. et al. BridgeData V2: a dataset for robot learning at scale. In Conference on Robot Learning (2023).
16. Sermanet, P. et al. Generating robot constitutions and benchmarks for semantic safety. Preprint at https://arxiv.org/abs/2503.08663 (2025).
17. Hawes, N. et al. The STRANDS project: long-term autonomy in everyday environments. IEEE Robot. Autom. Mag. 24, 146-156 (2017).
18. Wang, K. & Christensen, H. I. TritonBot: first lessons learned from deployment of a long-term autonomy tour guide robot. In IEEE RO-MAN (2018).
19. Wang, K., Javed, O. & Christensen, H. I. Robotic reliability engineering: experience from long-term TritonBot development. Preprint at https://arxiv.org/abs/1908.10495 (2019).
20. Valner, R. et al. Scalable and heterogeneous mobile robot fleet-based task automation in crowded hospital environments: a field test. Front. Robot. AI 9, 922835 (2022).
21. Bu, F., Fischer, K. & Ju, W. Making sense of robots in public spaces: a study of trash barrel robots. ACM Trans. Hum.-Robot Interact. (2025).
22. Braga, R. G., Tahir, M. O., Iordanova, I. & St-Onge, D. Robotic deployment on construction sites: considerations for safety and productivity impact. Preprint at https://arxiv.org/abs/2404.13143 (2024).
23. Herath, D. et al. Robots and aged care: a case study assessing implementation of service robots in an aged care home. In IEEE RO-MAN (2023).
24. Keith, R. & La, H. M. Review of autonomous mobile robots for the warehouse environment. Preprint at https://arxiv.org/abs/2406.08333 (2024).
25. Mankins, J. C. Technology Readiness Levels: a white paper. NASA Office of Space Access and Technology (1995).
26. Mitchell, M. et al. Model cards for model reporting. In Proc. FAT* (2019).
27. Gebru, T. et al. Datasheets for datasets. Commun. ACM 64, 86-92 (2021).
28. NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology (2023).
29. ISO/IEC 42001. Information technology - Artificial intelligence - Management system. International Organization for Standardization (2023).
30. ISO/IEC/IEEE 15026-2. Systems and software assurance - Assurance case. International Organization for Standardization (2022).
31. SAE International. J3016: Taxonomy and definitions for terms related to driving automation systems. SAE International (2021).
32. ISO 34503. Road vehicles - Test scenarios for automated driving systems - Specification for operational design domain. International Organization for Standardization (2023).
33. European Union. Regulation (EU) 2024/1689, Artificial Intelligence Act (2024).
34. ISO 10218-1. Robotics - Safety requirements - Part 1: Industrial robots. International Organization for Standardization (2025).
35. ISO/TS 15066. Robots and robotic devices - Collaborative robots. International Organization for Standardization (2016).
36. ISO 13482. Robots and robotic devices - Safety requirements for personal care robots. International Organization for Standardization (2014).
37. Ames, A. D. et al. Control barrier functions: theory and applications. In European Control Conference (2019).
38. Hsu, K.-C., Hu, H. & Fisac, J. F. The safety filter: a unified view of safety-critical control in autonomous systems. Annu. Rev. Control Robot. Auton. Syst. 7, 1-30 (2024).
39. Koopman, P. & Wagner, M. Autonomous vehicle safety: an interdisciplinary challenge. IEEE Intell. Transp. Syst. Mag. 9, 90-96 (2017).
40. Kirkpatrick, J. et al. Overcoming catastrophic forgetting in neural networks. Proc. Natl Acad. Sci. USA 114, 3521-3526 (2017).
41. Raji, I. D. et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In Proc. ACM FAccT (2020).
42. Boateng, G. O. et al. Best practices for developing and validating scales for health, social, and behavioral research. Front. Public Health 6, 149 (2018).
43. Messick, S. Validity of psychological assessment. Am. Psychol. 50, 741-749 (1995).
44. Campbell, D. T. & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol. Bull. 56, 81-105 (1959).
45. OECD and Joint Research Centre. Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD Publishing (2008).
46. Bainbridge, L. Ironies of automation. Automatica 19, 775-779 (1983).
47. Hart, S. G. & Staveland, L. E. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In Human Mental Workload 139-183 (North-Holland, 1988).
48. Endsley, M. R. Toward a theory of situation awareness in dynamic systems. Hum. Factors 37, 32-64 (1995).
49. Sheridan, T. B. Humans and Automation: System Design and Research Issues (Wiley, 2002).
50. Suchman, L. Plans and Situated Actions: The Problem of Human-Machine Communication (Cambridge Univ. Press, 1987).
51. Suchman, L. Making work visible. Commun. ACM 38, 56-64 (1995).
52. Star, S. L. & Strauss, A. Layers of silence, arenas of voice: the ecology of visible and invisible work. Comput. Support. Coop. Work 8, 9-30 (1999).
53. Elish, M. C. Moral crumple zones: cautionary tales in human-robot interaction. Engaging Sci. Technol. Soc. 5, 40-60 (2019).
54. Shafer, G. & Vovk, V. A tutorial on conformal prediction. J. Mach. Learn. Res. 9, 371-421 (2008).
55. Angelopoulos, A. N. & Bates, S. Conformal prediction: a gentle introduction. Found. Trends Mach. Learn. 16, 494-591 (2023).
56. Quigley, M. et al. ROS: an open-source Robot Operating System. In ICRA Workshop on Open Source Software (2009).

57. Macenski, S., Foote, T., Gerkey, B., Lalancette, C. & Woodall, W. Robot Operating System 2: design, architecture, and uses in the wild. Sci. Robot. 7, eabm6074 (2022).
58. Todorov, E., Erez, T. & Tassa, Y. MuJoCo: a physics engine for model-based control. In IEEE/RSJ International Conference on Intelligent Robots and Systems 5026-5033 (2012).
59. Makoviychuk, V. et al. Isaac Gym: high performance GPU based physics simulation for robot learning. Preprint at https://arxiv.org/abs/2108.10470 (2021).
60. McMahan, B. et al. Communication-efficient learning of deep networks from decentralized data. In Proc. AISTATS (2017).
61. Dwork, C. Differential privacy. In Proc. ICALP (2006).
62. Kairouz, P. et al. Advances and open problems in federated learning. Found. Trends Mach. Learn. 14, 1-210 (2021).
63. Rieke, N. et al. The future of digital health with federated learning. npj Digit. Med. 3, 119 (2020).
64. Dwork, C. & Roth, A. The Algorithmic Foundations of Differential Privacy. Found. Trends Theor. Comput. Sci. 9, 211-407 (2014).
65. Cox, D. R. Regression models and life-tables. J. R. Stat. Soc. Series B Stat. Methodol. 34, 187-220 (1972).
66. McCullagh, P. & Nelder, J. A. Generalized Linear Models, 2nd edn (Chapman and Hall, 1989).
67. Saaty, T. L. The Analytic Hierarchy Process (McGraw-Hill, 1980).
68. Billard, A. et al. A roadmap for AI in robotics. Preprint at https://arxiv.org/abs/2507.19975 (2025).