A Roadmap for Physical AI Foundation Models in Robotics

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

Abstract

Vision-language-action models excel at semantic generalization, yet real-world robotics imposes latency, jitter, energy, thermal and safety constraints that monolithic models cannot absorb. We argue that Physical AI foundation models must become cloud-edge-robot systems that separate time scales, safety authority and data lifecycles. We propose this architecture, an Action Proposal Contract and Physical Intelligence Metrics as a roadmap for assurance-ready robotics.

Keywords: Physical AI; robot foundation models; embodied intelligence; cloud-edge-robot architecture; safety assurance; digital twin; data flywheel; robotics evaluation

Introduction: from digital generalization to physical embodiment

Foundation models have reshaped digital AI by scaling representation learning across language, vision and multimodal data [1-3]. Robotics now faces a related but harder question: what does it mean for a model to be foundational when its outputs move bodies, contact objects and affect people? Vision-language-action (VLA) models, robot transformers, diffusion policies and language-conditioned planners have made this question concrete [4-14]. They show that semantic priors and broad robot experience can improve action generation, but they also reveal a category error: a robot action is not a sentence.

In the physical world, generalization matters only if it survives closed-loop control, communication delay, jitter, battery limits, heat, payload variation, calibration drift, human proximity and safety procedures. A high-level action may consume energy, occupy space, exert force and create irreversible damage. The point is not to reject VLA models. They are indispensable sources of semantic compression, cross-task priors and action proposals. The point is that they are insufficient as the sole unit of foundation in robotics.

We define a Physical AI foundation model as a reusable system foundation for embodied autonomy: one that combines semantic reasoning, perception, action proposals, verified skills, safety authority and real-world feedback across cloud, edge and robot layers. This Perspective makes three contributions. First, it reframes Physical AI foundation models as cloud-edge-robot systems rather than monolithic models. Second, it introduces the Action Proposal Contract as an assurance-ready interface between model output and robot execution. Third, it proposes Physical Intelligence Metrics for evaluating deployable autonomy beyond task success.

The missing dimension in current benchmarks

Current robot foundation model research can be read through five lenses: model-centric work on architectures and policies; data-centric work on demonstrations, teleoperation, human video and synthetic data; embodiment-centric work on transfer across arms, grippers, mobile bases, drones and humanoids; system-centric work on cloud, edge and robot allocation; and deployment-centric work on setup, safety cases, standards, data ownership, maintenance and return on investment.

The field is strongest in the first two lenses. RT-1, RT-2, OpenVLA, Octo, pi0, diffusion policies, action chunking, SayCan, PaLM-E and related systems have advanced semantic action generation and robot policy learning [4-14]. Open X-Embodiment, DROID, BridgeData V2, RoboNet and egocentric video resources have increased data scale and diversity [15-19]. Yet even the largest public robot datasets remain small compared with internet-scale language and vision corpora, and they remain biased toward tasks that can be collected safely and repeatedly.

Existing benchmarks are necessary but incomplete. Meta-World, ManiSkill, CALVIN, LIBERO, BEHAVIOR, RoboSuite, RLBench, Habitat and AI2-THOR measure multi-task manipulation, simulation generalization, long-horizon instruction following and embodied navigation [20-29]. They rarely make near misses, unsafe-instruction rejection, energy per successful task, communication-loss robustness, calibration effort or mean time between human interventions central outcomes. As a result, a policy can improve benchmark success while remaining unsuitable for a factory, hospital, public building or home.

This gap matters because robotics is embodied. The same action label, such as grasp, has different physical meanings for a suction gripper, parallel gripper, dexterous hand or humanoid arm. The same instruction can require different margins depending on payload, floor condition, human proximity, local regulation and emergency-stop availability. Foundation models therefore need interfaces to bodies and sites, not only interfaces to datasets.

Proposed framework: the cloud-edge-robot architecture

Cloud robotics and edge computing have long argued that computation can be distributed across networks and devices [30-35]. The Cloud-Edge-Robot (CER) framework proposed here is not simply another topology. It is an allocation principle: semantic generalization moves upward when latency tolerance and context breadth increase, while safety authority moves downward as physical risk and real-time criticality increase.

The first principle is time-scale separation. Cloud systems handle long-horizon reasoning, retrieval, large-scale simulation, fleet learning and model governance. Edge systems handle local perception, semantic maps, short-horizon planning, privacy filtering and runtime verification. Robot-embedded systems handle state estimation, skill execution, collision avoidance, force limits and emergency stopping. The second principle is safety-authority separation. VLA output should be treated as a proposal, not a command. Higher layers may propose intent, skills or target states, but verifiers decide whether the proposal is valid, and robot-local controllers retain final stop authority.

The third principle is data-lifecycle coupling. Deployment should generate structured logs of successes, failures, near misses, human interventions, energy use, latency, privacy state and maintenance events. These logs should feed digital-twin reconstruction, rare-case generation, retraining and regression testing. Physics simulators, world models and site digital twins are not replacements for real robots; they are tools for asking what-if questions and measuring reality gaps before redeployment [36-42].

Figure 1 summarizes this paradigm shift from monolithic VLA framing to a Cloud-Edge-Robot system foundation. Table 1 summarizes the allocation of primary functions, time-scale constraints and safety authority across the cloud, edge and robot layers.

[Insert Figure 1 about here]

[Insert Table 1 about here]

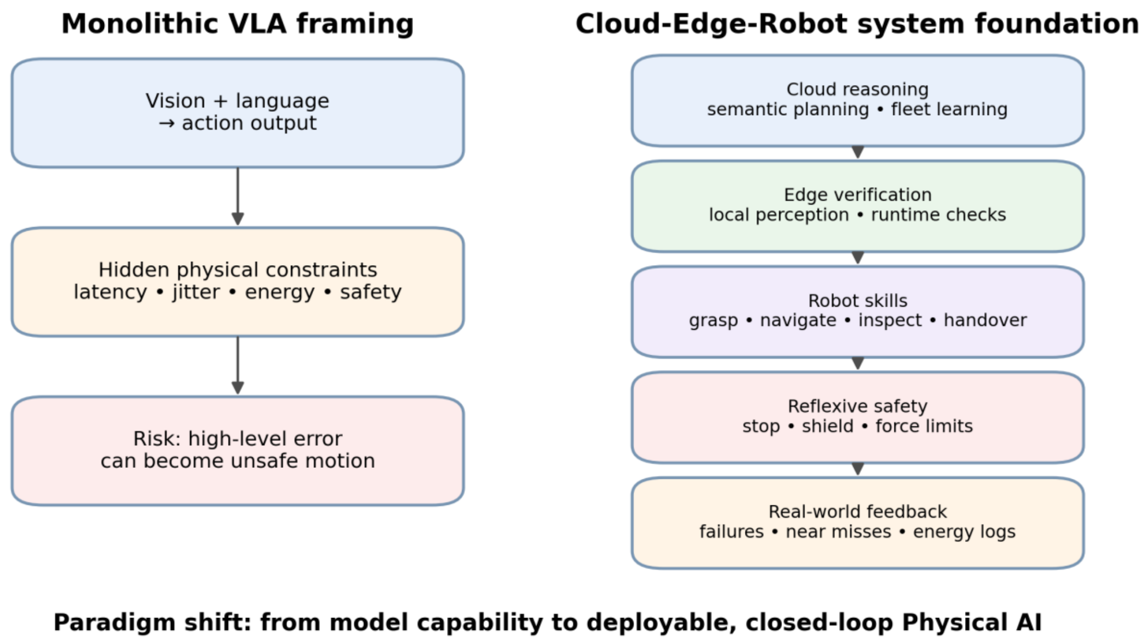

Figure 1 | The CER paradigm shift. Monolithic VLA framing treats model output as the primary unit. The proposed Cloud-Edge-Robot view separates cloud reasoning, edge verification, robot skills, reflexive safety and real-world feedback.

Table 1 | Function and constraint allocation across CER layers

Layer	Primary functions	Time scale and constraints	Safety authority
Cloud	Long-horizon reasoning, retrieval, fleet learning, simulation and model governance	Seconds to days; high compute acceptable when energy and carbon are accounted at system level	May propose plans but should not hold direct stop authority
Edge	Local VLA/VLM inference, semantic maps, short-horizon planning, runtime verification and privacy filtering	Milliseconds to seconds; constrained by site hardware, heat, connectivity and update procedures	Approves, rejects or modifies proposals before execution
Robot	Skills, state estimation, collision avoidance, actuator control and emergency stop	Microseconds to tens of milliseconds; constrained by battery, payload, cooling and certification	Retains final motion and stop authority

Action Proposal Contracts and closed-loop safety

A practical interface for CER is the Action Proposal Contract (APC). Instead of sending an unstructured action from a VLA model to a controller, the system should pass a structured proposal that is inspectable, rejectable, replayable and auditable. The APC does not claim that current VLA systems are safety-certified. It specifies the minimum information required to make model-generated behavior compatible with assurance engineering. Table 2 specifies the proposed Action Proposal Contract fields and illustrates how each field supports assurance-ready execution.

[Insert Table 2 about here]

Safety should then be layered. Semantic safety asks whether the intended outcome is dangerous, ambiguous, adversarial or socially inappropriate. Operational safety asks whether the action is allowed in this site, by this robot, at this time and under this procedure. Runtime assurance checks trajectories against constraints, shields unsafe actions and triggers fallback policies. Physical or reflexive safety implements emergency stops, protective stops, force limits, speed limits and actuator-level constraints. Control barrier functions, safety filters, shielding and runtime assurance provide relevant mechanisms, but VLA-controlled robots also require explicit links between AI safety and robotics safety [43-47]. Figure 2 illustrates how the safety stack filters model-generated proposals before execution and how deployment evidence feeds a real-world data flywheel for verified learning.

[Insert Figure 2 about here]

This link is also regulatory and organizational. Industrial, collaborative and mobile robot standards, functional-safety standards, fail-safe design guidance and AI risk-management frameworks provide essential anchors [48-56]. They do not, by themselves, solve semantic hazards, role-dependent permissions, prompt-level ambiguity or model-update safety cases. APCs can make these concerns visible at the boundary between model reasoning and physical execution.

Table 2 | Action Proposal Contract as an assurance-ready interface

APC field	Purpose	Example content
Task intent	States what the model is trying to achieve	Deliver object to specified human; inspect panel; open drawer
Skill identifier	Links semantic intent to a verified skill or skill family	grasp_v3; navigate_site_map; handover_safe
Target state	Defines the desired physical or symbolic end state	Object pose; inspection waypoint; recipient confirmation
Preconditions	Lists conditions that must be true before execution	Known object pose; free workspace; authorized user
Constraints	Makes limits explicit for verification	Speed, force, keep-out zone, payload, human proximity
Time-to-live	Prevents stale proposals from becoming motion	Expire after 500 ms or after state-estimator update
State confidence	Communicates perception and localization uncertainty	Object confidence, map confidence, tracking covariance
Authority level	Specifies whether the proposal can execute, be modified or must be reviewed	execute-if-verified; require-human-approval; reject
Fallback policy	Defines safe behavior when verification fails	stop; retreat; ask clarification; hand over to teleoperation
Audit identifier	Enables replay, incident analysis and regression testing	proposal ID, model version, site ID, log pointer

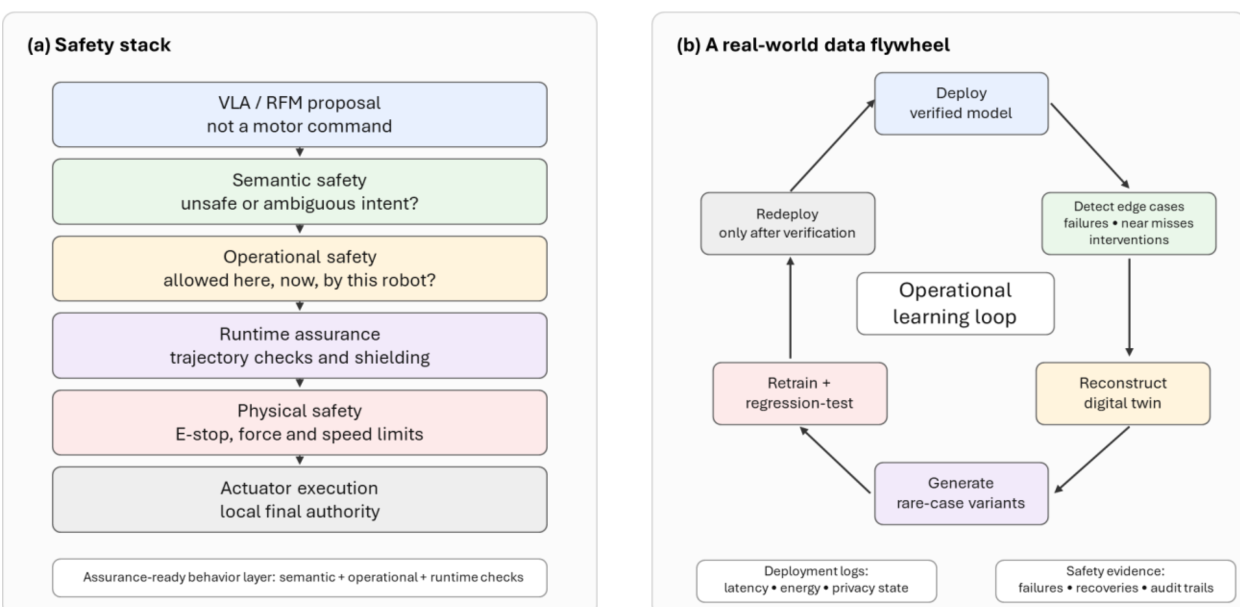

Figure 2 | Two-panel assurance loop. (a) A layered safety stack converts model output from proposal to permitted execution. (b) A real-world data flywheel converts deployments into verified learning through edge-case detection, digital-twin reconstruction, rare-case generation, retraining and regression testing.

Physical Intelligence Metrics and the 2030 roadmap

Task success remains necessary, but it is not sufficient. A robot that completes a task slowly, unsafely, with excessive energy, frequent recalibration or brittle connectivity is not physically intelligent in an operational sense. Physical Intelligence Metrics should therefore extend task success toward six dimensions: Adaptability, Safety Resilience, Real-Time Robustness, Energy Efficiency, Embodiment-Aware Transfer and Maintainability.

Table 3 defines the six Physical Intelligence Metrics and provides example measurement protocols for each dimension.

[Insert Table 3 about here]

These metrics should be use-case weighted rather than collapsed into a universal score. Industrial manipulation emphasizes safety and timing. AMR logistics emphasizes throughput, energy and intervention rate. Domestic robotics emphasizes privacy, low power and false refusal. Healthcare emphasizes semantic, operational and physical safety. Disaster response emphasizes connectivity-loss robustness and recovery. Energy and carbon accounting should be explicit because embodied intelligence includes computation, motion and infrastructure, not only model accuracy [57-60].

The 2030 horizon is not a prediction of full general-purpose autonomy. It is a practical planning horizon for testing whether foundation-model-based robotics can move from demonstrations to assurance-ready deployment in constrained domains. By 2026, the question is whether VLA/RFM policies can support reliable closed-loop operation in bounded settings. By 2028, on-device or edge-deployed VLA systems, digital-twin validation, fleet learning and early semantic-safety evaluation should become testable industrial practices. By 2030, the community should aim for architectures in which behavior proposals are structured, safety authority is explicit, model updates are regression-tested and physical intelligence is measured beyond task success.

Figure 3 summarizes the proposed roadmap toward 2030 and links the transition from model emergence to assurance-ready systems with the six Physical Intelligence Metrics.

[Insert Figure 3 about here]

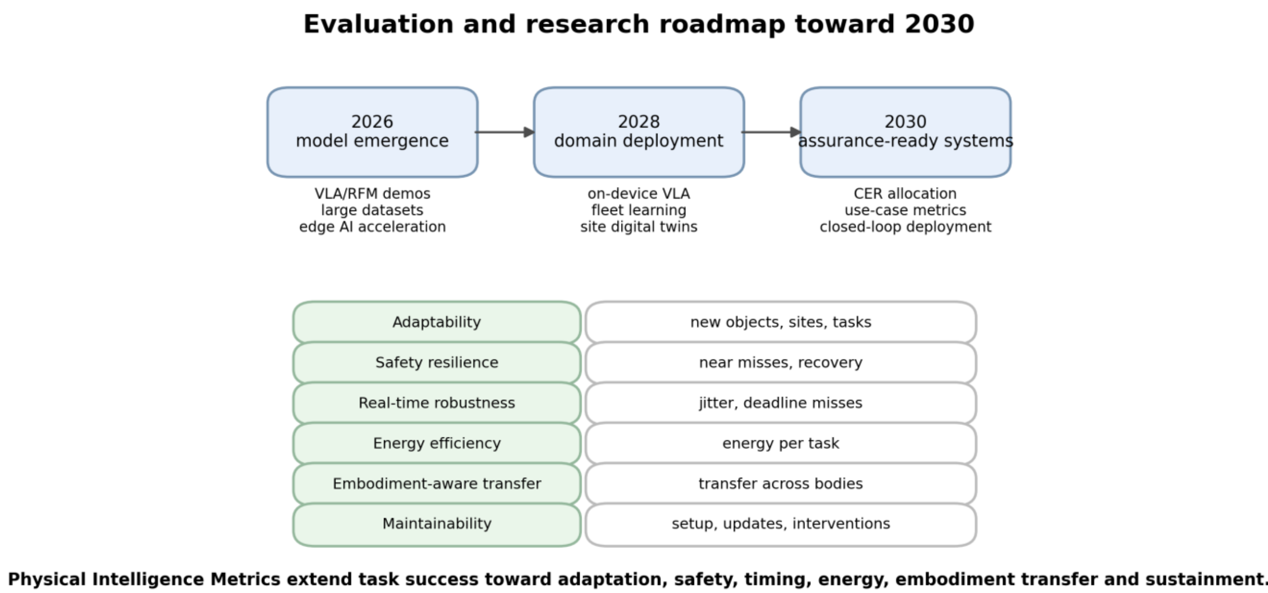

Figure 3 | Evaluation and research roadmap toward 2030. The roadmap moves from model emergence to domain deployment and assurance-ready systems, while the metrics extend task success toward adaptation, safety, timing, energy, embodiment transfer and sustainment.

Table 3 | Physical Intelligence Metrics for deployable robotics foundation models

Metric	Definition	Example measurement protocol
Adaptability	Maintains performance under new objects, scenes, tasks or sites	Seen/unseen task split; few-shot improvement after k demonstrations; site-to-site transfer
Safety Resilience	Rejects unsafe behavior and recovers from faults	Near-miss rate; unsafe-instruction rejection; false refusal; recovery time; fallback success
Real-Time Robustness	Meets timing requirements under load, jitter and disconnection	Deadline misses; jitter sensitivity; packet-loss injection; safe degradation
Energy Efficiency	Uses embodied and computational energy economically	Energy per successful task; compute per action; robot/edge/cloud energy accounting
Embodiment-Aware Transfer	Transfers across bodies while respecting morphology-specific constraints	Cross-gripper, arm, mobile-base or humanoid transfer with body-specific safety envelopes
Maintainability	Can be installed, calibrated, updated and sustained	Setup time; calibration hours; update regression tests; intervention rate; maintenance burden

Research agenda for assurance-ready Physical AI

This roadmap implies concrete priorities. Model researchers should study interfaces between VLA proposals and verified skill execution, not only end-to-end action generation. Dataset builders should include failures, recoveries, near misses, latency, energy and governance metadata. Simulation researchers should connect synthetic rare cases to real deployment logs and quantify reality gaps. Safety researchers should integrate semantic, operational, runtime and reflexive safety. System researchers should evaluate CER allocation under latency, disconnection, power and update constraints. Standards communities should begin defining audit logs, model-update safety cases and benchmark protocols for AI-generated robot behavior.

Table 4 translates this roadmap into priority research actions for model researchers, dataset builders, simulation researchers, safety researchers, system researchers and standards communities.

[Insert Table 4 about here]

Table 4 | Priority research agenda for assurance-ready Physical AI

Community	Priority toward 2030	Why it matters
Model researchers	Connect VLA proposals to verified skill execution	Prevents semantic capability from bypassing physical constraints
Dataset builders	Log failures, recoveries, near misses, energy, latency and governance metadata	Makes deployment learning measurable rather than anecdotal
Simulation researchers	Tie synthetic rare cases to real logs and quantify reality gaps	Turns digital twins into verification tools, not visual replicas
Safety researchers	Integrate semantic, operational, runtime and reflexive safety	Connects AI risk to robot motion and site procedures
System researchers	Evaluate CER allocation under power, disconnection, update and thermal constraints	Makes foundation models deployable on real hardware
Standards communities	Define audit logs, model-update safety cases and benchmark protocols	Creates shared evidence for safe scaling across sites

Conclusion and call to action

Physical AI foundation models should not be reduced to larger VLA models. Robotics requires foundations that are physically grounded, embodied, edge-deployable, safety-assured and continuously improved through real-world feedback. The unit of analysis should shift from the model to the closed-loop system.

The Cloud-Edge-Robot architecture is a call to allocate intelligence, authority and learning by time scale and risk. The Action Proposal Contract is a call to make model outputs inspectable and rejectable before they become motion. Physical Intelligence Metrics are a call to evaluate robots by adaptation, safety, real-time robustness, energy, embodiment-aware transfer and maintainability. If foundation models are to transform robotics by 2030, they must become foundations for safe and sustainable autonomy in the physical world.

Author contributions

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

Competing interests

The author declares no competing interests.

Funding

The author received no specific funding for this work.

References

1. Vaswani, A. et al. Attention is all you need. In Advances in Neural Information Processing Systems 30, 5998-6008 (2017).

2. Brown, T. B. et al. Language models are few-shot learners. In Advances in Neural Information Processing Systems 33, 1877-1901 (2020).

3. Bommasani, R. et al. On the opportunities and risks of foundation models. arXiv:2108.07258 (2021).

4. Brohan, A. et al. RT-1: Robotics Transformer for real-world control at scale. In Proceedings of Robotics: Science and Systems (2023).

5. Zitkovich, B. et al. RT-2: Vision-Language-Action models transfer web knowledge to robotic control. In Proceedings of The 7th Conference on Robot Learning, PMLR 229, 2165–2183 (2023).

6. Open X-Embodiment Collaboration. Open X-Embodiment: Robotic learning datasets and RT-X models. In Proceedings of the IEEE International Conference on Robotics and Automation, 6892–6903 (2024). https://doi.org/10.1109/ICRA57147.2024.10611477

7. Kim, M. J. et al. OpenVLA: An open-source vision-language-action model. In Proceedings of The 8th Conference on Robot Learning, PMLR 270, 2679–2713 (2025).

8. Octo Model Team et al. Octo: An open-source generalist robot policy. In Proceedings of Robotics: Science and Systems (2024). https://doi.org/10.15607/RSS.2024.XX.090

9. Black, K. et al. pi0: A vision-language-action flow model for general robot control. arXiv:2410.24164 (2024).

10. Chi, C. et al. Diffusion Policy: Visuomotor policy learning via action diffusion. The International Journal of Robotics Research 44, 1051–1083 (2025).

11. Zhao, T. Z., Kumar, V., Levine, S. & Finn, C. Learning fine-grained bimanual manipulation with low-cost hardware. In Proceedings of Robotics: Science and Systems (2023). https://doi.org/10.15607/RSS.2023.XIX.016

12. Ahn, M. et al. Do as I can, not as I say: Grounding language in robotic affordances. In Proceedings of The 6th Conference on Robot Learning, PMLR 205, 287–318 (2023).

13. Driess, D. et al. PaLM-E: An embodied multimodal language model. In Proceedings of the 40th International Conference on Machine Learning, PMLR 202, 8469–8488 (2023).

14. Liang, J. et al. Code as policies: Language model programs for embodied control. In Proceedings of the IEEE International Conference on Robotics and Automation (2023).

15. Khazatsky, A. et al. DROID: A large-scale in-the-wild robot manipulation dataset. In Proceedings of Robotics: Science and Systems (2024).

16. Walke, H. et al. BridgeData V2: A dataset for robot learning at scale. In Proceedings of The 7th Conference on Robot Learning, PMLR 229, 1723–1736 (2023).

17. Dasari, S. et al. RoboNet: Large-scale multi-robot learning. In Proceedings of the Conference on Robot Learning, PMLR 100, 885–897 (2020).

18. Ebert, F. et al. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. In Proceedings of Robotics: Science and Systems (2022).

19. Grauman, K. et al. Ego4D: Around the world in 3,000 hours of egocentric video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18995-19012 (2022).

20. Yu, T. et al. Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning. In Proceedings of the Conference on Robot Learning, PMLR 100, 1094–1100 (2020).

21. Mu, Y. et al. ManiSkill2: A unified benchmark for generalizable manipulation skills. In Proceedings of the International Conference on Learning Representations (2023).

22. Mees, O., Hermann, L., Rosete-Beas, E. & Burgard, W. CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. IEEE Robotics and Automation Letters 7, 7327-7334 (2022).

23. Liu, B. et al. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In Advances in Neural Information Processing Systems, Datasets and Benchmarks Track (2023).

24. Li, C. et al. BEHAVIOR-1K: A benchmark for embodied AI with 1,000 everyday activities and realistic simulation. In Proceedings of The 6th Conference on Robot Learning, PMLR 205, 80–93 (2023).

25. Zhu, Y. et al. RoboSuite: A modular simulation framework and benchmark for robot learning. arXiv:2009.12293 (2020).

26. James, S., Ma, Z., Arrojo, D. R. & Davison, A. J. RLBench: The robot learning benchmark and learning environment. IEEE Robotics and Automation Letters 5, 3019-3026 (2020).

27. Savva, M. et al. Habitat: A platform for embodied AI research. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 9339-9347 (2019).

28. Kolve, E. et al. AI2-THOR: An interactive 3D environment for visual AI. arXiv:1712.05474 (2017).

29. Shridhar, M., Manuelli, L. & Fox, D. CLIPort: What and where pathways for robotic manipulation. In Proceedings of The 5th Conference on Robot Learning, PMLR 164, 894–906 (2022).

30. Kuffner, J. J. Cloud-enabled robots. In Proceedings of the IEEE-RAS International Conference on Humanoid Robots (2010).

31. Kehoe, B., Patil, S., Abbeel, P. & Goldberg, K. A survey of research on cloud robotics and automation. IEEE Transactions on Automation Science and Engineering 12, 398-409 (2015).

32. Shi, W., Cao, J., Zhang, Q., Li, Y. & Xu, L. Edge computing: Vision and challenges. IEEE Internet of Things Journal 3, 637-646 (2016).

33. Satyanarayanan, M. The emergence of edge computing. Computer 50, 30-39 (2017).

34. Mao, Y., You, C., Zhang, J., Huang, K. & Letaief, K. B. A survey on mobile edge computing: The communication perspective. IEEE Communications Surveys & Tutorials 19, 2322-2358 (2017).

35. Chen, X., Jiao, L., Li, W. & Fu, X. Efficient multi-user computation offloading for mobile-edge cloud computing. IEEE/ACM Transactions on Networking 24, 2795-2808 (2016).

36. Todorov, E., Erez, T. & Tassa, Y. MuJoCo: A physics engine for model-based control. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 5026-5033 (2012).

37. Makoviychuk, V. et al. Isaac Gym: High performance GPU-based physics simulation for robot learning. In Advances in Neural Information Processing Systems, Datasets and Benchmarks Track (2021).

38. Coumans, E. & Bai, Y. PyBullet: a Python module for physics simulation for games, robotics and machine learning. Software at https://pybullet.org (2016).

39. Tobin, J. et al. Domain randomization for transferring deep neural networks from simulation to the real world. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 23-30 (2017).

40. Peng, X. B., Andrychowicz, M., Zaremba, W. & Abbeel, P. Sim-to-real transfer of robotic control with dynamics randomization. In Proceedings of the IEEE International Conference on Robotics and Automation, 3803-3810 (2018).

41. Kadian, A. et al. Sim2Real predictivity: Does evaluation in simulation predict real-world performance? IEEE Robotics and Automation Letters 5, 6670-6677 (2020).

42. Moyne, J. et al. A requirements driven digital twin framework: specification and opportunities. *IEEE Access* **8**, 107781–107801 (2020)

43. Ames, A. D., Xu, X., Grizzle, J. W. & Tabuada, P. Control barrier function based quadratic programs for safety critical systems. IEEE Transactions on Automatic Control 62, 3861-3876 (2017).

44. Fisac, J. F. et al. A general safety framework for learning-based control in uncertain robotic systems. IEEE Transactions on Automatic Control 64, 2737-2752 (2019).

45. Alshiekh, M. et al. Safe reinforcement learning via shielding. In Proceedings of the AAAI Conference on Artificial Intelligence 32 (2018).

46. Wabersich, K. P., Hewing, L., Carron, A. & Zeilinger, M. N. Probabilistic model predictive safety certification for learning-based control. IEEE Transactions on Automatic Control 67, 176-188 (2022).

47. Hsu, K., Hu, H. & Fisac, J. F. The safety filter: A unified view of safety-critical control in autonomous systems. Annual Review of Control, Robotics, and Autonomous Systems 7, 47–72 (2024).

48. International Organization for Standardization. ISO 10218-1:2025 Robots and robotic devices - Safety requirements for industrial robots - Part 1: Robots (2025).

49. International Organization for Standardization. ISO 10218-2:2025 Robots and robotic devices - Safety requirements for industrial robots - Part 2: Robot systems and integration (2025).

50. International Organization for Standardization. ISO/TS 15066:2016 Robots and robotic devices - Collaborative robots (2016).

51. International Electrotechnical Commission. IEC 61508: Functional safety of electrical/electronic/programmable electronic safety-related systems (2010).

52. International Organization for Standardization. ISO 13849-1:2023 Safety of machinery - Safety-related parts of control systems - Part 1: General principles for design (2023).

53. IEEE Standards Association. IEEE 7009-2024: IEEE standard for fail-safe design of autonomous and semi-autonomous systems (2024).

54. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1 (2023).

55. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) (2024).

56. 3rd Generation Partnership Project. TS 22.261: Service requirements for the 5G system (2024).

57. Popovski, P. et al. Wireless access for ultra-reliable low-latency communication: Principles and building blocks. IEEE Network 32, 16-23 (2018).

58. Strubell, E., Ganesh, A. & McCallum, A. Energy and policy considerations for deep learning in NLP. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, 3645-3650 (2019).

59. Schwartz, R., Dodge, J., Smith, N. A. & Etzioni, O. Green AI. Communications of the ACM 63, 54-63 (2020).

60. Patterson, D. et al. Carbon emissions and large neural network training. arXiv:2104.10350 (2021).

61. Lacoste, A. et al. Quantifying the carbon emissions of machine learning. arXiv:1910.09700 (2019).