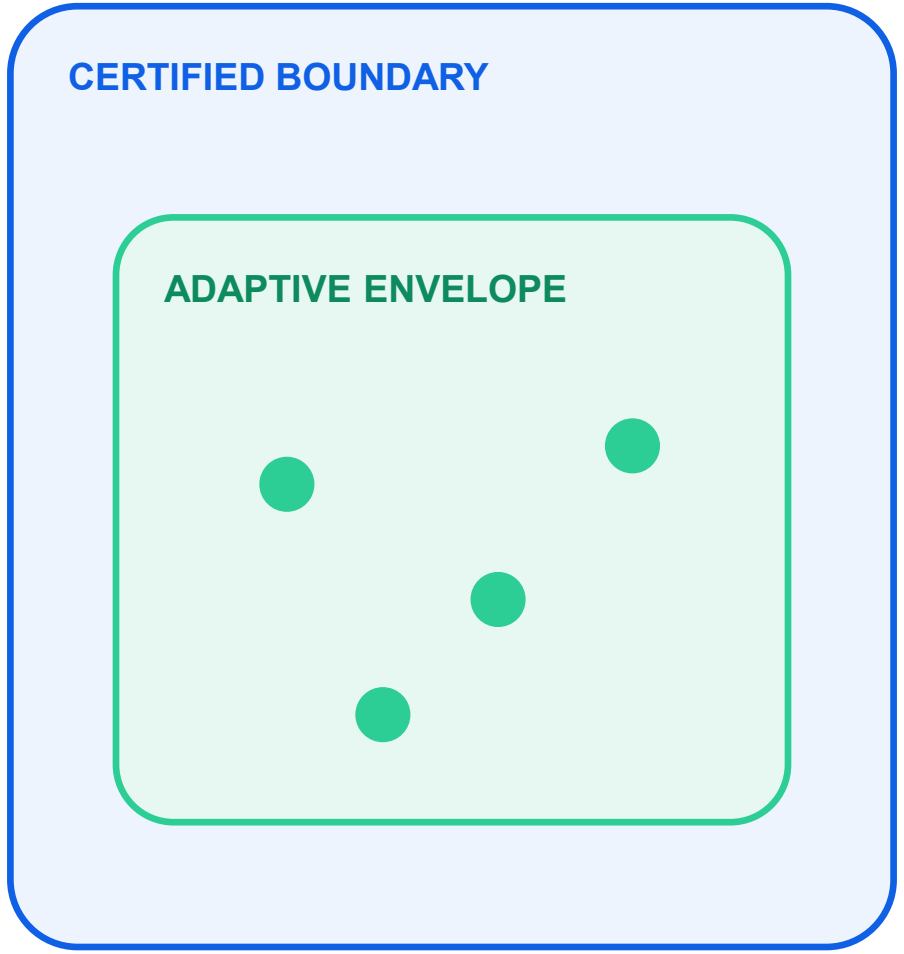
Bounded, Not Just Benchmarked

Rethinking evaluation for embodied AI in the real world

The bottleneck isn’t only capability. It’s **evaluation design, and who we measure.**



Yonah Welker, Visiting Technologist

ITU Workshop: Embodied AI and Multimedia Technology. Session 1: Core Technologies. Hangzhou, 16 Oct 2026.

One ecosystem, one set of rules

The ecosystem

140+

Embodied public and assistive AI systems in the OECD repository

- Exoskeletons
- Social robotics
- Bionic limbs
- Smart wheelchairs

What it runs on

INFRASTRUCTURE AND CAPACITIES

- Edge devices, low latency
- Compressed models
- Energy-efficient architectures
- Durable, mature chips
- Multimodal sensor fusion

Often with low or no connectivity

Bounded and benchmarked

SHARED RULES

1. Say who the data represents
2. Fixed safety boundary, adaptive envelope, runtime check
3. Test across populations

Deployable. Comparable. Trusted.

Evidence from real deployments flows back into the rules

A score without a “who”

Most benchmarks ask whether a system completes the task. Fewer ask who it was tested on.

39 of 827

human-robot interaction studies directly reported participants' disability status (ACM/IEEE HRI, 2006 to 2022)

Seaborn, Barbareschi and Chandra, Int. J. of Social Robotics, 2023

36.3% → 23.7%

word error rate on Parkinson's speech: model trained on typical speech vs. with Parkinson's recordings

Speech Accessibility Project, JSLHR, 2024

58%

of people with facial differences had trouble being recognised by passport e-gates

Face Equality International survey, 2024

One aggregate score hides who the system fails.

The typical-body assumption hides in four places

Data collection	Simulation	Test site	Evaluation
TYPICAL ASSUMPTION Coverage targets name space, task and environment, not who is in the data	**TYPICAL ASSUMPTION** One reachability value per object; fidelity ignores body differences (proportions, gait)	**TYPICAL ASSUMPTION** One nominal passage geometry; test subjects left unspecified	**TYPICAL ASSUMPTION** Verification data does not say who appears in it
PROPOSED CHANGE Name embodiment and human differences (body proportions, gait patterns) as a coverage dimension	**PROPOSED CHANGE** Reachability by agent class; human and embodiment differences (proportions, gait) as an 8th fidelity dimension	**PROPOSED CHANGE** Accessible-mobility clearance profile (ISO 21542); documented test population	**PROPOSED CHANGE** Disclose coverage; include at least one assistive-device case
WG2: TS-mda	**WG2: TS-psa3d, TS-pdse**	**WG8: TS-tcts §7.3, §8.5**	**WG8: TS-wm §7.2, §12.1**

Reachability, passages and clearances are spatial intelligence questions, and the answer depends on who is moving through the space.

1. Say who the data represents

Comparable

A number is only comparable if you know who is behind it.

Same space, different answer. Can this cabinet handle be reached by:

- a bipedal humanoid?
- a wheelchair user?
- a prosthetic hand?

Dataset coverage statement

Provenance	where and how captured; real or synthetic
Embodiment	body sizes, mobility aids, prosthetics, atypical gait
Communication	speech, sign, AAC, eye gaze
Annotators	who labelled the data, and how
Known gaps	stated, not hidden

Connects to WG2, WG8, F.BA-EAI

2. Separate safety guarantees from adaptive behavior

Deployable

CERTIFIED SAFETY BOUNDARY (fixed)

max speed, minimum clearance, stopping distance

ADAPTIVE ENVELOPE (inside the boundary)

routing, speed modulation, grip and gait modulation

Function	Fixed (certified)	Adaptive (inside)
Speed control	max speed	runtime speed cap
Path planning	no-go zones	route selection
Prosthetic grip	max force and torque	grip modulation by user and task

RUNTIME CHECK

Validates the plan before adaptation takes effect

Fallback: degraded mode, fleet-assisted, safe stop or operator

Public and human-interfaced embodied systems frequently run at the edge, on compressed, energy-efficient architectures, so fallback has to work with low or no connectivity.

Speed limits can be certified, fleet-level learning cannot.

OECD.AI: Physical AI governance

3. Test across populations as standard practice

Trusted

	Typical body	Wheelchair, mobility aid	Prosthetic, assistive device	Atypical gait, posture	Speech, sign, AAC
Passage and turning (TS-tcts §7.3)	○	○	○	○	○
Human-EAI safety (TS-tcts §8.5)	○	○	○	○	○
Subject consistency (TS-wm §12.1)	○	○	○	○	○
Verification data (TS-wm §7.2)	○	○	○	○	○

Three changes

Document who the test subjects are

Add an accessible-mobility clearance profile (ISO 21542)

Report results by group, not only in aggregate

No result in a cell means no claim for that group.

Where this is already being worked on in FG-EAI

Group	Documents	Proposal	Stage
WG2	TS-mda, TS-psa3d, TS-pdse, TS-mdmp	Embodiment coverage, reachability by agent class, 8th fidelity dimension, coverage disclosure	Proposed
WG8	TS-tcts, TS-wm, F.BA-EAI	Clearance profile, test subjects with different bodies and gaits, coverage disclosure, one shared cross-reference	Proposed
WG5	TS-coeai §7.3, §8.1	Constrained-connectivity profile, bounded adaptation formalized	Proposed
WG7 / TG7.1	TS-Acc-interacts, TR-Acc-up	Bounded adaptation for oversight; disability-disaggregated evaluation, worked use cases	Proposed
TG6.4	New Technical Report	Disability-centred AI in healthcare: data, certification, evaluation gaps	Proposed

These are proposals for discussion, not adopted text. Experts are welcome in all of these groups.

Three asks

1. **Report the “who” (trained on, tested on) with every benchmark number** — Comparable
2. **Separate fixed safety claims from adaptive ones, and check at runtime** — Deployable
3. **Put atypical bodies, embodiments and aids in the test plan, not the appendix** — Trusted

Deployable. Comparable. Trusted. Not just benchmarked.

References

Evidence cited

Celi, L. A., et al. (2022). Sources of bias in artificial intelligence that perpetuate healthcare disparities: A global review. *PLOS Digital Health*, 1(3).
Hasegawa-Johnson, M., et al. (2024). Community-supported shared infrastructure in support of speech accessibility. *Journal of Speech, Language, and Hearing Research*.
Northumbria University, Centre for Responsible AI, and Face Equality International (2024). Survey of people with facial differences and facial recognition.

Policy and guidance

OECD (2023). *Using AI to support people with disability in the labour market: Opportunities and challenges*. OECD Publishing.
WHO (2024). *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models*. World Health Organization.
Welker, Y. *Sovereign AI for assistive and public technologies: Critical digital capacities*. OECD.AI.
Welker, Y. *Physical AI: Implications for governance, standards and testing frameworks*. OECD.AI.

Standards

ETSI EN 301 549 V3.2.1 (2021). *Accessibility requirements for ICT products and services*.
IEC 62061:2021. *Safety of machinery: Functional safety of safety-related control systems*.
IEC TS 62998-1:2019. *Safety of machinery: Safety-related sensors used for the protection of persons*.
ISO 10218-1:2025 and ISO 10218-2:2025. *Robotics: Safety requirements for industrial robots*.
ISO 21542:2021. *Building construction: Accessibility and usability of the built environment*.
ISO/IEC Guide 71:2014. *Guide for addressing accessibility in standards*.