Robotic Data

WEBINAR 01 — ATTENDEE RESOURCE

Physical AI Data Readiness Checklist

Why Physical AI Needs More Than Synthetic Data

Closing the gap between simulation and real-world performance.

Synthetic and simulated data are valuable, but they cannot reproduce every unexpected condition, human interaction, environmental change, or long-tail event. This resource shows how real-world data complements simulation — and helps you assess whether your data is ready for the real world.

Duration: 45 minutes • Dates to be announced

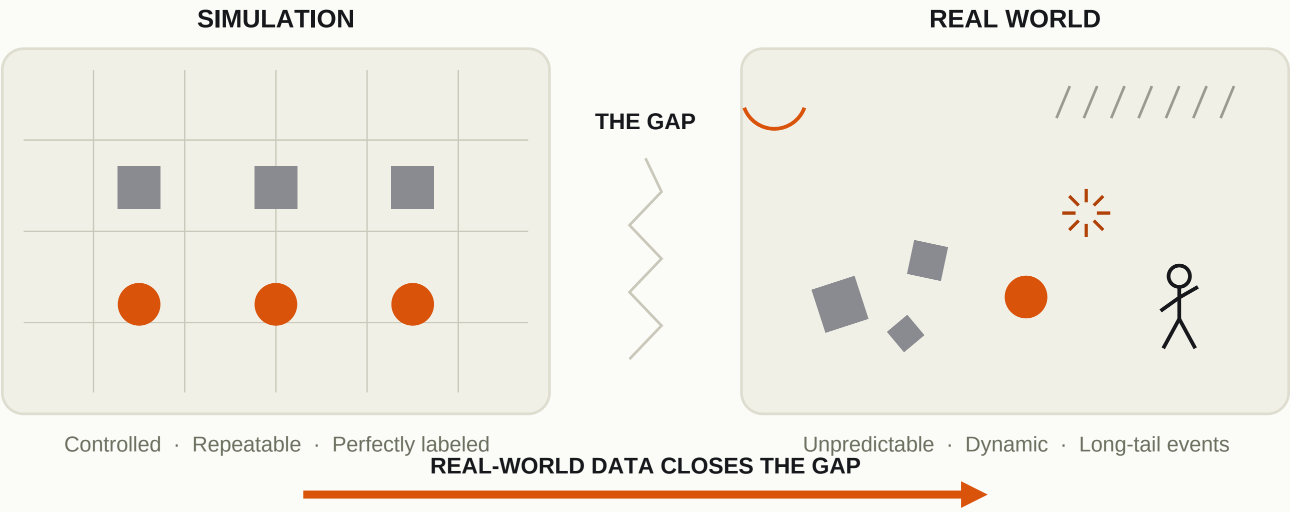
Robotic Data — Data for the Physical AI Revolution

OVERVIEW

Why Physical AI needs more than synthetic data

Simulation and synthetic data are powerful: they scale cheaply, generate perfect labels, and let you rehearse dangerous scenarios safely. But a model that has only seen a rendered world has never seen the real one — with its sensor noise, weather, changing spaces, and people who do not follow scripts.

The sim-to-real gap



The distance between rehearsal and reality is the sim-to-real gap, and it is where Physical AI systems fail in deployment. The gap cannot be rendered away; it is closed with real-world data — captured in the environments, conditions, and human contexts where the system will actually operate.

> **The core question**
>
> This resource ends with a readiness checklist. The question it answers: is your data — not your model — the thing limiting real-world performance?

SECTION 01

01 Physical AI versus traditional digital AI

Digital AI reads, writes, and classifies; its mistakes cost a retry. Physical AI moves through the world; its mistakes have momentum. That difference changes what data must capture.

	Traditional digital AI	Physical AI
Where it acts	Screens, documents, databases	Streets, warehouses, homes, skies
Cost of error	A wrong answer — retry is free	A collision, a drop, an injury — errors are physical
Environment	Static or slowly changing	Dynamic: weather, lighting, layout, wear, seasons
Inputs	Clean, structured, curated	Noisy multi-sensor streams (camera, LiDAR, force, audio)
Time	Mostly stateless requests	Continuous perception–action loops; timing matters
People	Users behind an interface	Humans sharing physical space, behaving unpredictably

Implication for data

Training data for Physical AI must capture embodiment: real sensor statistics, real physics, real timing, and real human behavior. Text scraped from the internet built digital AI; the physical world has no equivalent corpus — it has to be collected.

SECTION 02

02 Understanding the real-world data gap

Simulators reproduce what their authors thought to model. The gap is everything nobody thought to model — and it is systematic, not random. The conditions simulation most often misses:

Sensor reality

- True noise profiles, motion blur, rolling-shutter artifacts, lens flare, and glare — rendered sensors are too clean.
- Degradation over time: dirty lenses, misaligned mounts, dropped frames, calibration drift.

Environmental change

- Weather interacting with sensing: rain on LiDAR, fog, low sun angles, wet reflective surfaces.
- Spaces that change: construction, rearranged layouts, seasonal foliage, new signage — the world your map described last month is not this month's world.

Human behavior

- People cut corners, hesitate, gesture, carry awkward objects, and break rules — behavior that scripted agents rarely reproduce.
- Interaction effects: people respond to the robot itself, creating feedback loops no simulator anticipates.

Long-tail events

- The events that dominate deployment risk are, by definition, rare: near-misses, unusual object combinations, one-in-ten-thousand scenes.
- Long-tail events cannot be enumerated in advance — they must be encountered, captured, and fed back.

> **Why the gap is systematic**
>
> Simulation error is not random noise around reality — it is biased toward the clean, the average, and the expected. Models inherit that bias and are most confident exactly where they are most wrong.

SECTION 03

03 Why controlled datasets fail in uncontrolled environments

Even real data can fail if it was collected under lab conditions. A dataset gathered in one facility, one season, one demographic, or one scripted routine teaches the model that those conditions are the world.

- **Distribution shift.** Deployment conditions drift away from collection conditions; accuracy decays silently until failures cluster.
- **Lab bias.** Uniform lighting, tidy scenes, and cooperative demonstrators produce models that stall on clutter, glare, and haste.
- **Spatial overfitting.** A robot trained in one building learns its quirks as physics; a second site is effectively a new planet.
- **Scripted behavior.** Instructed participants move predictably; real people do not. Models trained on scripts misread genuine intent.

What uncontrolled-ready data looks like

- Collected across sites, seasons, times of day, and demographics — diversity as a design requirement, not an accident.
- Includes failures, occlusions, degraded sensors, and interrupted tasks — the messy episodes controlled programs discard.
- Refreshed on a cadence matched to how fast the environment changes.

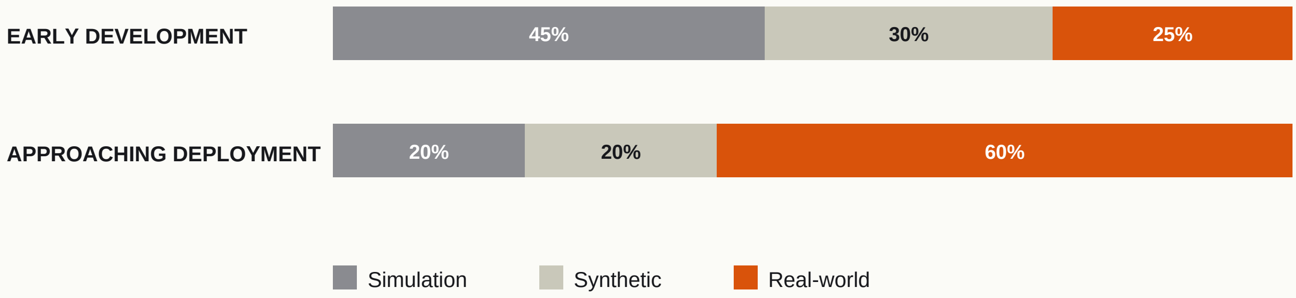
SECTION 04

04 Real data, synthetic data, and simulation

This is not a contest — mature programs use all three, each where it is strongest, and shift the balance as they approach deployment.

Simulation	Explore cheaply and safely: dangerous scenarios, control tuning, millions of rehearsal hours, perfect labels.
Synthetic data	Amplify variation: rebalance rare classes, sweep lighting and texture, stress-test perception at low cost.
Real-world data	Anchor and validate: true sensor statistics, genuine human behavior, spatial context, the long tail — and the only trustworthy measure of readiness.

Rebalancing toward reality



Early on, simulation dominates because iteration speed matters most. As deployment approaches, real-world data must take over — for training on the long tail and, above all, for honest evaluation. Illustrative proportions; the right mix depends on your task and risk profile.

SECTION 05

05 The Physical AI Data Readiness Checklist

Work through the six categories below. Check a box only if you could defend it with evidence. Score yourself at the end.

1 — Task and operating envelope

- [ ] The task and its success criteria are written down and measurable
- [ ] The operating envelope is defined: environments, objects, people, and explicit out-of-scope conditions
- [ ] Failure modes are enumerated with severity ratings

2 — Data inventory

- [ ] The current mix of real, synthetic, and simulated data is quantified
- [ ] Provenance and licensing are known for every dataset in training
- [ ] Sensor configurations in the data match deployment hardware

3 — Real-world coverage

- [ ] Lighting, weather, and seasonal variation are represented
- [ ] Data spans multiple sites and regions, not just the lab or pilot site
- [ ] Long-tail and edge events are deliberately collected and tracked
- [ ] Sensor noise, degradation, and occlusion appear in training data

4 — Human factors

- [ ] Unscripted, natural human behavior appears in the data
- [ ] Human–robot interaction scenarios are covered, including responses to the robot itself
- [ ] The people in the data reflect the diversity of people in deployment

5 — Spatial context and change

- [ ] Spatial context (maps, geometry, semantics) is current and validated
- [ ] Environmental change is tracked: construction, layout shifts, seasonal effects
- [ ] A data-recency policy exists for environments that change

6 — Performance diagnosis

- [ ] The model is evaluated on held-out real-world data, not just simulation
- [ ] The sim-to-real performance gap is measured and tracked over time
- [ ] Failures are clustered and attributed: data gap vs. model vs. control
- [ ] Plateau analysis shows whether more of the current data still helps
- [ ] A feedback loop routes deployment failures into targeted collection

Score your readiness

17–21 checked: your data program is deployment-grade — keep the loop turning. **10–16:** real gaps exist; prioritize coverage and diagnosis items first. **Below 10:** data, not the model, is very likely what limits your real-world performance.

SECTION 06

06 Live audience Q&A

The session closes with live Q&A. The most useful questions name a specific task, environment, or failure mode you are facing. Note yours here.

Webinar 01 — Physical AI Data Readiness

45 minutes • Dates to be announced • www.roboticdata.com — Robotic Data creates the real-world datasets that power autonomous vehicles, humanoid robots, drones, and the next generation of Physical AI.