■ REPORT · V1.0

CLASSIFICATION / PUBLIC

FORESIGHT FUTURE INSTITUTE · INVESTIGATION

Ground Truth

What Was Accepted?

Who accepts that the worlds machines learn and are tested in are adequate, for what use — and can anyone outside see the decision?

> *Acceptance is specific to a use. Public records let us trace one military simulation accreditation from version to use, limits and acceptor; for the two self-driving approvals we examined, we could not recover that link.*

REPORT · V1.0 · EVIDENCE CUTOFF 11 NOV 2026 · PUBLISHED 12 NOV 2026

CONTENTS

Ground Truth

Lock	How to read this issue	5
Ch 00	Three worlds and one sentence	7
Ch 01	Who builds the worlds	8
Ch 02	Military first?	11
Ch 03	War as a test range	14
Ch 04	Does the world predict reality?	16
Ch 05	The precedent: when a simulation counts as evidence	18
Ch 06	Rules and their exclusions	20
Ch 07	The mandate	23
Horizon	Next issue	27
Register	Appendix A. Case register: search notes	28
Register	Appendix B. Actor ledger (27 actors)	29
Register	Appendix C. Interpretation Lock, summary	31
Register	Appendix D. Evidence inventory (dated)	32
Register	Sources	34
Register	Glossary	36

How to read the figures. Each figure carries its own legend. Gold marks published quantitative evidence or a role named in a rule; red marks acceptance or disclosure that is missing, excluded or unresolved. Colour never encodes a measured quantity of safety.

How to read the numbers. Company figures are their own reports and are marked [M] or [V]; funding is capital supply, not buyer demand; a correlation is association, not calibration. Evidence labels in notes: [M] maker record · [V] vendor, assessor or investor with an interest · [G] government or standards body · [R] research · [P] press or secondary · NR not resolved.

Acceptance is specific to a use. Public records let us trace one military simulation accreditation from version to use, limits and acceptor; for the two self-driving approvals we examined, we could not recover that link.

Machines that drive, fly, walk and fight now learn and are tested in worlds that someone built: physics simulators, digital twins, generative world models and the toolchains around them. This issue asks who accepts that such a world is adequate, for what, on what evidence — and whether anyone outside can see that decision. Acceptance routes exist and are deliberately specific to a use. What we audited is the public trace: for each identified case, which links between

a world version, its evidence and an acceptance decision public records supply, and which they leave unresolved. The one complete autonomy-adjacent record we recovered is military, and it took two official reports, published years apart, to assemble.

What we found

1. **Acceptance routes exist.** US military rules require a simulation's accreditation to record the version, the intended use and the accepting authority (DoDI 5000.61). EU type approval for driverless vehicles assesses the credibility of the simulation toolchain, including version control and validation thresholds (Regulation 2022/1426); German law requires simulators to show "no significant difference" from real trials (AFGBV); a new maritime code lets authorities accept simulation results under an approved verification plan. Our dated inventory (Appendix D) holds 24 predictive or comparative entries on simulated and physical results, seven further physical-outcome entries and five benchmark diagnostics; entries are not independent studies.
2. **They accept a world for a use — by design.** Verification and validation exist to establish that a model is adequate for a stated decision and conditions. A public claim that a system is "validated" does not tell an outsider which decision, evidence or limits were accepted.
3. **The public trace varies by case.** The Pentagon's test office reported that the F-35 programme's executive committee "accredited JSE, with limitations, for IOT&E on September 1st, 2023"; its FY2025 report identifies the F-35 cockpit configuration used in that test as 30S02. Read together, the two reports link a simulation version, a use, limits and an acceptor — for a simulation that tests a human-piloted aircraft, and at the level of the aircraft model rather than the whole environment. For the two self-driving approvals we examined, approvals and roles are public; the accepted simulator version and its evidence were not retrieved. Our three comparison cases — an FDA tool qualification, an NRC licence amendment and a 2017 US Army training-simulator accreditation — have the same kind of public summary.
4. **Fidelity figures depend on the evaluation design.** A robot benchmark whose simulated scores had an average correlation of 0.92 with real results in its authors' evaluation had 0.40 in another team's evaluation of different policies on seven of its tasks. These are different evaluations, not one estimate of a simulator's quality.
5. **EU robot rules arrive in two steps.** Under the enacted Machinery Regulation, a notified-body route applies from 20 January 2027 to specified safety components with self-evolving, machine-learning behaviour; additional AI requirements must apply through delegated acts by 2 August 2028. The enacted text we reviewed does not establish a public register of training-world acceptance; we did not locate a published draft or adopted delegated act in the records searched through the cutoff (11 November 2026). The machinery and AI regimes have different military-related exclusions (chapter 6).
6. **Evidence production can be vertically integrated.** Simulation suppliers sit inside or beside the companies whose systems they test — a US defence vendor building a Pentagon evaluation platform, an AI-pilot maker that bought a simulation business, a Chinese robot maker evaluating in its own simulator. Supplying the tool does not by itself confer the role of acceptor. It does make the public trace more important.

Why now. On 4 September 2026 the UN expert group on lethal autonomous weapons adopted elements saying states should understand a system's capabilities and limits by "conducting testing and evaluation, including adequately realistic simulations" in the expected operational environment. The CCW Review Conference is scheduled to consider them on 16–20 November. The text keeps state responsibility; it sets no common technical threshold for "adequately realistic" and names no technical accreditation body.

Version 1.0, published 12 November 2026, after three adversarial rounds, two dated pre-cutoff sweeps (23 October and 7 November) and a final check to the evidence cutoff, 11 November 2026, 23:59 GST. Independent, unfunded; AI assistance and review disclosed in How to read this issue.

INTERPRETATION LOCK

How to read this issue

This issue tests hypotheses and reports where each stands; none is assumed. The rules below were first written on 23 October 2026, after a first evidence pass, and revised after each of three adversarial rounds. Some clarifications were made after the case results were known; they are logged as such (Appendix C). A pre-cutoff sweep on 23 October recovered a record that changed one case's coding and defeated H1d as first stated; that change is logged too (Amendment A4). The rules are not preregistered.

The question. Who accepts that a world a model learns or is tested in is adequate — for what use, on what evidence — before the model drives, flies, walks or fights? And can anyone outside trace that acceptance?

What we mean by "world". The object of acceptance is a model or toolchain at a stated version: a physics simulator, a digital twin, a generative world model, a synthetic-sensor pipeline. Real-world data, including combat data, is an input used to build, train or calibrate such worlds; access to data is not itself a world.

Hypotheses, as they stand.

	CLAIM	STATUS AT THE CUTOFF, 11 NOVEMBER (UNCHANGED SINCE 23 OCTOBER)
H1a	Few actors that make physical-use claims publish quantitative, version-specific evidence that their worlds predict physical outcomes for a material claimed task	Not resolved: eligibility and the material-task rule are not yet frozen (chapter 1)
H1b	Verification capacity grows more slowly than world-building	Not resolved: we did not locate matched capacity and workload data
H1c	No identifiable party accepts a world's adequacy	**Rejected in its general form** (chapter 5). Replaced by H1d
H1d	In the identified autonomy cases reviewed here, we did not recover a complete public link between simulation version, validation evidence and acceptance decision	**Rejected as stated** on 23 October (Amendment A4): the F-35 JSE accreditation has a complete public link at summary level across two DOT&E reports (chapter 7). Retained in narrowed form for the two self-driving approvals in the register (C2, C3). Not a prevalence estimate
H2	Military demand leads physical autonomy and its training worlds; civilian uses follow	Not resolved, on money and on rules (chapter 2)

Evidence ladder. Each claim of the form "world X is adequate to train or test system Y" is rated on its own, not by company: **L0** demonstration; **L1** internal metric claimed; **L2-T** quantitative physical results after training in a simulated world; **L2-P** quantitative comparison of simulated and physical task outcomes or trajectories for the same policies or system; **L3** independent replication or scoped third-party test, with commissioning disclosed; **L4** accepted in a named safety case by a regulator, certifier or military test authority. *Not resolved* (NR) is kept apart from L0. Physical results from planning with a world model learned from real data, and comparisons of generated video with filmed physical experiments, are recorded separately. Provenance and independence are recorded separately: an academic paper may still evaluate its authors' own benchmark.

Boundaries.

- "Not found in the records we searched" is not "not done". Leading developers and authorities may hold complete records privately.
- Acceptance for a stated use is normal assurance practice, not a failure.
- A role named in a rule is not proof that every programme complied.
- Correlation is not calibration: a simulator can order policies correctly and still overstate how often they succeed.
- Funding rounds, valuations, contract ceilings and investors are not evidence of who buys.
- Allegations and accidents are not proof of a simulator defect.
- This is analysis of markets, rules and verification. It contains no operational, targeting or design detail on any weapon system.
- For Gulf states we record only measures the state or its entities announced, with dates and without evaluation.

What would change our conclusions. The tests, with thresholds, are set out in Appendix C. For example: if a frozen, eligible set of actors met a strict disclosure standard at 80% or more, H1a would be rejected; if civilian buyers accounted for more than 60% of paid demand for training and test environments even after unknown amounts were counted as military, H2's demand claim would be rejected. H1d is rejected for any eligible case in the register for which a public record, published by the cutoff, links an identified simulation version, its intended use, a substantive evidence or limitations summary and the accepting body; one complete eligible case defeats the claim as stated, and any such amendment will be logged. That happened for C1 on 23 October. A version is an explicit release or build, an evaluated-method revision, or a named and dated frozen configuration baseline; a decision date alone does not count.

Labels. [M] maker's own record · [V] vendor, assessor or investor with an interest · [G] government, military or standards body · [R] research · [P] press or secondary · NR not resolved. Labels describe where information came from, not how reliable or independent it is.

How this issue was made. Research and drafting by the Foresight Future Institute with Claude (Anthropic); a separate model acted as research partner and as opponent in three adversarial rounds and two pre-cutoff evidence sweeps. Its findings narrowed the thesis twice, and the 23 October sweep overturned our coding of the military case; every disposition is logged, including the findings we did not adopt. Anthropic and OpenAI are not subjects of this issue; their investors and partners include companies named here.

CHAPTER 00

Three worlds and one sentence

Three public accounts from 2026 illustrate different parts of the assurance question: a world built to test a car, a simulator bought by the maker of an AI pilot, and a physical trial of humanoids.

A road that never existed. On 6 February Waymo announced a world model built on Google DeepMind's Genie 3 that generates camera and lidar scenes its cars have never driven: rare hazards, counterfactual manoeuvres, dashcam footage turned into simulation. The announcement shows side-by-side videos; in it and in an August post on lessons from more than 200 million fully autonomous miles, we found no quantitative measure of how closely this particular world matches the road [M]. Waymo publishes a safety-case method and describes calibration and an AI critic run across simulated miles. The gap is specific: in the records we inspected, we could not see how this world version was validated, or by whom it was accepted.

A simulator bought by the pilot's maker. On 22 June Shield AI completed its acquisition of Aechelon, a maker of high-fidelity visual simulation and sensor models whose products support the Pentagon's Joint Simulation Environment. Its chief executive said every action "must be tested and validated in simulation and fed back into a continuous, closed data loop" [M]. The company that builds an AI pilot now owns a simulation business of the kind such pilots are tested in. Who accepts a given simulation for a given use is not stated in the announcement.

A field at Quantico. On 9–10 September the US Army ran physical demonstrations for its xTechHumanoid competition: 103 submissions, five winners, an aim to "establish a technical baseline" [G]. These were tests on real ground. The Army's account does not say whether entrants trained in simulation, and describes no standard for accepting simulation evidence; its silence does not show that none applies.

The sentence. On 4 September, in Geneva, the UN Group of Governmental Experts on lethal autonomous weapons adopted its final report. Among the elements it asks the Seventh Review Conference (16–20 November) to consider and endorse, states should understand a system's capabilities and limits by "conducting testing and evaluation, including adequately realistic simulations" in its expected operational environment (§38(a)). The same text keeps states and individuals responsible under international law and says "accountability cannot be transferred to machines" (§§40–41) [G].

The text says what testing is for. It does not set a common threshold for *adequately realistic*, name a technical accreditation body, or say how anyone outside a state's own system would see that the test was met. In civilian sectors, those decisions are already made by test agencies, type-approval authorities and notified bodies — one world, one version and one use at a time.

CHAPTER 01

Who builds the worlds

Among the 27 civilian and dual-use world builders and robot-model makers listed here, seven selected claims have a cited comparison of simulated and physical task outcomes or trajectories. Three more actors — Galbot, Odyssey and Fourier — report quantitative physical outcomes after training in simulation; Meta's physical-planning result is recorded separately. No L4 record was located for the exact claims coded. These are counts of our table, not a test of H1a; the ledger is in Appendix B.

The money, with its status. In 2026: World Labs, founded by Fei-Fei Li, announced $1 billion (18 February; no valuation disclosed) [M], and on 28 September AMD announced a definitive agreement to acquire it in an all-stock deal valued at about $8.2 billion, expected to close by the end of 2026 [M]; AMD's quarterly report for the quarter ended 26 September 2026 still lists it among pending acquisitions [M]. AMI Labs, co-founded by Yann LeCun, raised a $1.03 billion seed at a $3.5 billion pre-money valuation (announced 10 March) [P]; Skild AI announced $1.4 billion at over $14 billion (14 January) [M]; Wayve announced a closed $1.2 billion Series D with up to $300 million more conditional (25 February) and a $60 million extension (15 April) [M]; Waabi announced $750 million closed and $250 million contingent (28 January) [M]. Emulate, founded by former members of DeepMind's Genie team, was reported on 17 September to be in talks for up to $700 million; October reporting describes investments at different prices [P]. We did not verify a completed round or amount and exclude the proposed total. Physical Intelligence was reported by Bloomberg on 27 March to be in talks to raise about $1 billion at more than $11 billion; later press reports describe a completed raise of about $1 billion in May [P], and a Fidelity fund's holdings report dated 31 August 2026 lists Physical Intelligence Series C shares [V]. We could not open the press reports in full, the filing gives neither the round's amount nor its closing, and there is no company confirmation, so we do not count the $1 billion as raised. Bloomberg reported on 27 October that the company was in talks for new funding at a $20 billion valuation [P]; not counted.

The evidence, claim by claim. The table shows the strongest public evidence we found for each actor's world. It rates evidence, not companies: a firm can hold a strong study for one task and only demonstrations for its product.

ACTOR	WHAT ITS WORLD DOES	STRONGEST PUBLIC EVIDENCE (DATE)	LEVEL
World Labs	Rebuilds real scenes as simulations to train and rank robot policies	Bimanual cube handover: 2,000 simulated and 100 real trials per checkpoint; rankings preserved; chart only, no scalar metric (28 Jul 2026)	L2-P
AgiBot (China)	Genie Sim 3.0 simulation, synthetic demonstrations, automated evaluation	Four tasks × four training regimes for one base policy = 16 cells, 50 real and 50 simulated trials per cell; R^2 = 0.931 (not Pearson r), slope about 1.02, no interval (v4, 14 Aug 2026)	L2-P/T
Physical Intelligence (co-author)	PolaRiS evaluation of robot foundation models	Six paired scenes at two institutions, four policies, 20 real and 50 simulated rollouts each, after simulation co-training; mean r = 0.90, worst scene 0.81 (Dec 2025)	L2-P
1X	World model to evaluate humanoid policies	Checkpoint ranking against blind real A/B runs; held-out success/failure prediction accuracy 63% → 71% (Jun 2025)	L2-P
Waabi	Waabi World driving simulator	Maker-defined trajectory realism score of 99.7%; scenario count and interval not disclosed (Mar 2025)	L2-P

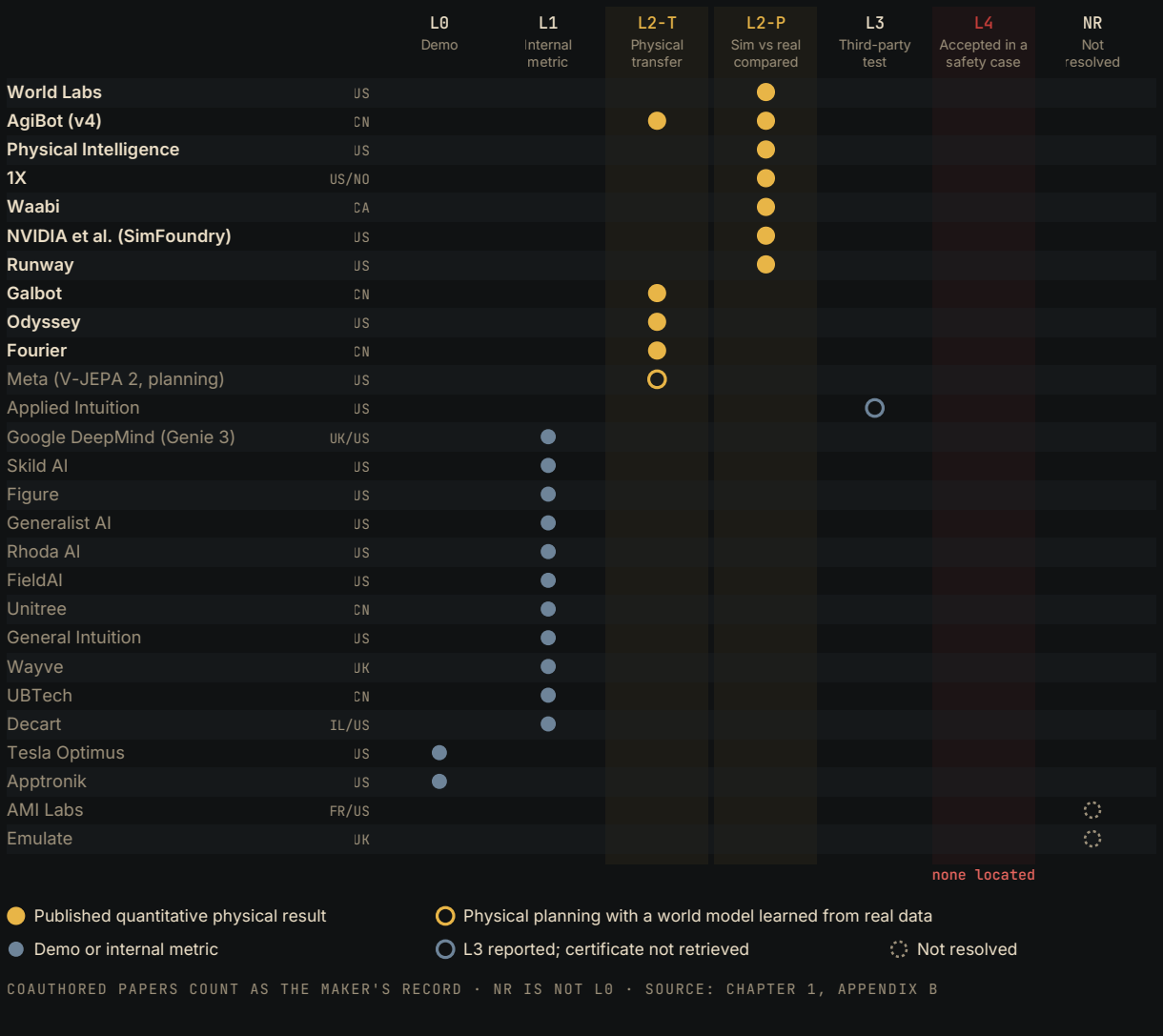
ACTOR	WHAT ITS WORLD DOES	STRONGEST PUBLIC EVIDENCE (DATE)	LEVEL
NVIDIA and co-authors	SimFoundry automated scene generation for policy evaluation	Seven manipulation tasks, five policy architectures; mean Pearson r = 0.911 (v4, Aug 2026)	L2-P
Runway	GWM-Robotics world model for policy evaluation	Eight policies on one arm; 1,450 simulated rollouts compared with RoboArena physical results; Pearson r = 0.95 for policy ranking, MMRV 0.033; physical trial count and interval not given (27 Feb 2026)	L2-P
Galbot (China)	Synthetic data for grasping	300 physical trials per method; success allows three attempts (May 2025)	L2-T
Odyssey	Generative world model used to train driving policies	Policies trained only in simulation drove about 77% as far between safety-driver interventions on real roads as policies trained on real footage; trial count and interval not given (15 Sep 2026)	L2-T reported
Fourier (China)	Humanoid policies trained in NVIDIA Isaac Gym	89% physical floor-to-stand success on the GR-2 robot after simulation training; trial count not given (NVIDIA account, 6 Nov 2024) [M][V]	L2-T reported
Meta	V-JEPA 2 latent world model, learned from real video and robot data, used for planning	Physical planning in two labs, success out of ten trials with initial conditions varied across trials (Jun 2025)	Physical planning (recorded separately)
Applied Intuition	Simulation and validation tools for vehicles and defence	Reports TÜV SÜD certification of its automotive virtual-validation platform; certificate not retrieved (19 Aug 2026)	L3 reported; certificate not retrieved
Google DeepMind	Genie 3 interactive worlds	Consistency and interaction claims; no physical comparison (5 Aug 2025). A separate Google result, physical success rising from 56% to 75% after training in a MuJoCo simulation twin (13 Oct 2026), does not test Genie 3	L1
Skild AI, Figure, Generalist, Rhoda, FieldAI, Unitree, General Intuition, Wayve, UBTech, Decart	Robot models, humanoids, driving and generative worlds	Internal metrics, specifications or deployment claims; no paired physical comparison (Appendix B names each metric)	L1
Tesla Optimus, Apptronik	Humanoids	Demonstrations	L0
AMI Labs, Emulate	World models	No released physical-transfer item found	NR

Sources: company and paper links in the table; dated sources for grouped rows, the 27-actor matrix and the financing ledger are in the evidence appendix. Company-coauthored papers count as the maker's own record [M]. Waymo, discussed in chapter 0, is treated as a deployment case, not a world vendor.

— FIGURE 01 · EVIDENCE LADDER: ONE SELECTED CLAIM PER ACTOR

Seven of 27 world builders publish a sim-versus-real task comparison

One selected claim per actor, checked 23 Oct 2026 · categories, not a cumulative scale or a ranking · no public L4 record located



COAUTHORED PAPERS COUNT AS THE MAKER'S RECORD · NR IS NOT L0 · SOURCE: CHAPTER 1, APPENDIX B

What the table shows. Quantitative evidence exists, from US, Canadian, Norwegian-American and Chinese makers, for manipulation, humanoid locomotion and road driving. It is narrow: a few tasks, a few policies, sometimes after the policies were adapted to the simulator. World Labs states the design goal plainly: a useful simulation "need not match real-world success rates exactly"; it must support the same decisions as reality. That is a defensible engineering aim for the policy-selection task studied; the result does not by itself establish overall robot safety.

What it does not show. For none of the claims coded did we locate a public L4 — an identified regulator, certifier or military test authority accepting a named world version in a named safety case. That may happen privately. Chapter 5 shows where rules already require such acceptance, and chapter 7 what a public trace would contain.

CHAPTER 02

Military first?

The Cold War pattern — military money first, civilian products as by-products — cannot be confirmed or rejected for this field with the records we assembled. It is unresolved for money and, on closer inspection, for the rules too.

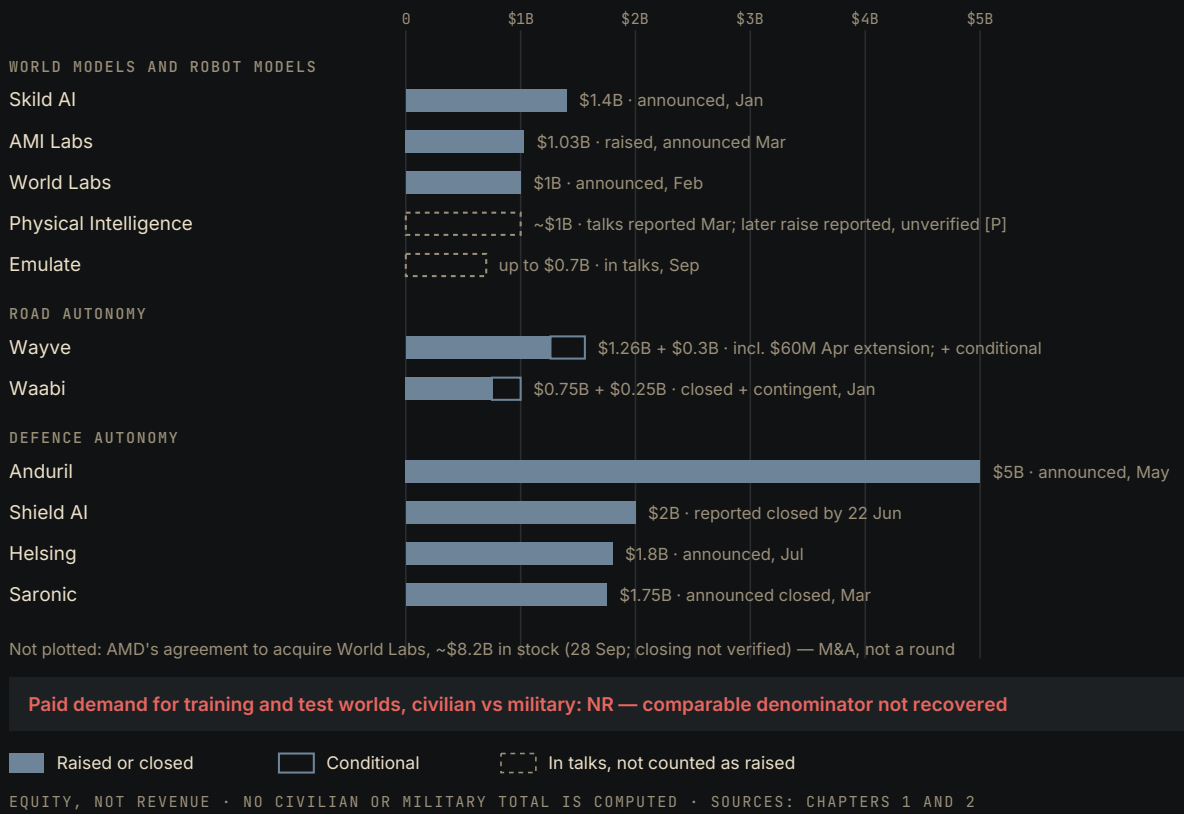
The case for. Defence-autonomy companies announced large private rounds in 2026: Anduril $5 billion at $61 billion (13 May) [M]; Shield AI $2 billion at $12.7 billion post-money, including $500 million of preferred equity, reported as closed in its 22 June acquisition announcement [M]; Helsing $1.8 billion at $18 billion (13 July) [M]; and the uncrewed-vessel maker Saronic $1.75 billion at a stated $9.25 billion, announced closed on 31 March [M]. Simulation capability is moving inside these companies: Shield AI acquired Aechelon; the Pentagon's Chief Digital and AI Office selected Applied Intuition on 14 May to build the platform for its Autonomy Factory programme, intended for programme offices, developers and operators to "develop, validate and certify autonomous systems" [M]. IQT, the independent not-for-profit investor established by the CIA, is listed among the investors of Skild AI and of Odyssey, which names defence among the industries its world models could transform [M].

The case against. The cited financings also include companies describing civilian applications: World Labs (spatial worlds for creativity, robotics and science; now agreed to be acquired by a chipmaker), AMI Labs (industry, robotics, healthcare, wearables), Wayve and Waabi (road transport, with Uber as strategic partner). General Intuition grew out of a gaming-clip platform; Runway out of creative video. The quantitative sim-to-real evidence in chapter 1 is civilian. Even CSIS's study of Russian drone autonomy stresses commercial and dual-use origins of the components and models involved.

The Gulf and Israel: announced measures. These entries record what states or their entities announced, with dates.

DATE	STATE	ANNOUNCED MEASURE
21 Jan 2026	UAE	At UMEX and SimTEX in Abu Dhabi, the Tawazun Council announced Ministry of Defence contracts worth AED 2.36 billion over two days, including development of a NIMR vehicle by EDGE (AED 10.6 million), described by Gulf News as autonomous, and development of QX uncrewed aerial systems (AED 45.1 million) (WAM; Gulf News) [G][P]
22 Jan 2026	Qatar	DIMDEX closed with more than 70 agreements, contracts and MoUs worth over QAR 18.5 billion; exhibits included unmanned and remote-controlled systems (QNA) [G]
10 Feb 2026	Saudi Arabia	At the World Defense Show, the Transport General Authority and SAMI Autonomous Company, a subsidiary of PIF-owned SAMI, signed an MoU on regulations and technical frameworks for autonomous vessels (SPA) [G]
11 Feb 2026	Qatar	Qatar Investment Authority announced participation in Apptronik's Series A extension; it had announced joining Applied Intuition's Series F on 17 June 2025 (QIA, Apptronik; QIA, Applied Intuition) [G]
8 Sep 2026	Israel	The IDF Chief of Staff announced a new Unmanned Systems and Artificial Intelligence Branch and directed that it be inaugurated during the coming Hanukkah, which begins in early December (Breaking Defense, 9 Sep [P]; IDF account partner-checked [G])

— FIGURE 02 · SELECTED 2026 FINANCINGS, WITH TRANSACTION STATUS



Why the money cannot decide it. These figures mix financing, valuations, a proposed acquisition and procurement announcements; they do not measure comparable paid demand for simulation. Military-first is a claim about demand — who pays for training and test environments, and who deploys first. A $249 million blanket purchase agreement for autonomy test software is a ceiling shared among several awardees, not money spent [M]. An investor is not a buyer. The records we reviewed do not yield a comparable split of paid demand for simulation between civilian and military customers, so we draw no civilian-versus-military spending total.

Why the rules cannot decide it either. The US military has an explicit, current framework for accepting simulations as evidence. DoD Instruction 5000.61 (current version effective 17 September 2024) defines accreditation as "the official certification that a model, simulation, or distributed simulation is acceptable for use" for a specific purpose, and requires a record of the version, intended use, criteria, data sources and "identification of the user or accreditation authority and record of the accreditation decision" [G]. DoD Manual 5000.102 (9 December 2024) applies this to digital twins and other models used as test evidence: operational test agencies accredit models for operational test, live-fire test organisations for live-fire test, and DOT&E approves the strategy for programmes under its oversight; models "must be accredited for the intended use prior to the M&S runs for record" [G]. Such decisions are sometimes reported publicly: the most complete public acceptance record in our register is military (chapter 7). One case does not show which sector discloses more. Civilian sectoral precedents are older than the DoD-wide instruments cited here: US aviation regulators were approving flight simulators by the mid-to-late 1960s, and expanded simulator-use rules in 1980 (FAA, 2002); the US Atomic Energy Commission set binding requirements for reactor safety evaluation models on 4

January 1974 (NRC history, §I.B) [G]. The DoD-wide simulation directive cited here dates to 1994 and the VV&A instruction to 1996; those dates do not mark the beginning of military simulation assurance, and these precedents do not establish priority for matched autonomy capabilities. For autonomy specifically, a civilian maritime code adopted in 2026 provides for simulation results to be "reviewed and accepted" under an approved verification plan (chapter 5). Which sector writes binding validation rules first for the same capability has not been tested; our matched-pair test (H-8) remains open.

CHAPTER 03

War as a test range

The war in Ukraine has made combat imagery a source of training and evaluation data for autonomy, and Ukraine's government documents this in a dated public record. What the record shows is controlled data access with quality controls for perception models; it does not show a simulated world built or accepted from that data.

Ukraine: the data, in sequence.

DATE	MEASURE	WHAT THE RECORD ESTABLISHES
21 Jan 2026	Ministry of Defence launches Bravel Dataroom	A secure environment for developers to train military AI on relevant data [G]
12 Mar 2026	Government opens battlefield data to partners	"Millions of annotated frames" from "tens of thousands of combat flights"; models trained on an MoD platform without direct access to the databases; used for target-recognition networks in the DELTA system [G]
11 Jun 2026	More than 100 Ukrainian companies use the Bravel Dataroom	Government-reported adoption count [G]
10 Aug 2026	Avengers Labs opened to Ukrainian defence companies	Five million annotated frames for model development [G]
24 Aug and 25 Sep 2026	UK becomes first international partner of Avengers Labs; UK competition on drone swarming using its data	Imagery from daylight and infrared sensors capturing "millions of objects" [G]
Oct 2026	UK competition questions and answers	Focus is perception, "for training and evaluating perception models"; labels "standardised and versioned" and "reviewed and validated by a human"; data stays in a secured sandbox; assessment covers deployability and assurance, not algorithmic performance alone [G]

— FIGURE 03 · UKRAINE: DOCUMENTED DATA, TRAINING AND EVALUATION CAPABILITIES

Official accounts describe training and evaluation capabilities

Ukraine MoD and UK announcements, Jan–Oct 2026 · no model-specific acceptance record located · perception models



A simulator or generative world built or calibrated from this data: not shown in the sources reviewed

Documented in an official announcement · Not located · Sequence, not a traced link

SEPARATE PLATFORMS; DOCUMENTED CAPABILITIES, NOT ONE MODEL'S LINEAGE · NO OPERATIONAL DETAIL

The minister, Mykhailo Fedorov, put the purpose plainly: “The future of warfare belongs to autonomous systems.” The records describe versioned labels, human review and a platform capability for evaluation. They name Ukraine’s Ministry of Defence as the use authority; we did not recover a model-specific acceptance decision, or a link from these data to an accepted simulator or generative world. Frames, detections and training runs are different units, recorded here on the platform where each was reported.

Russia: assessed, not confirmed. An open-source study by CSIS (Kateryna Bondar, 13 April 2026), based on triangulated public and closed-channel sources and interviews, describes “functional independence at the tactical edge” rather than comprehensive autonomy. It reports applications built on existing open-weight models, one manufacturer’s vision models retrained on users’ own data, and a state training database; it assesses that one drone type has “likely” been fielded as fully autonomous, relying in part on Ukrainian technical analysis of recovered hardware; and it notes that a draft Russian AI law circulated in March 2026 would explicitly exempt defence and security applications [R]. The report also stresses commercial and dual-use origins of components and models.

China: procurement and statements. Reuters has reported on PLA procurement notices, university research and defence-industry products involving robots and AI (27 October 2025); a later Reuters review of more than 100 documents found tenders for humanoids and training technology, systems to collect sensor and terrain data, and simulations of robot assault teams, but no evidence of an armed humanoid with an operational unit (7 September 2026, via Asharq Al-Awsat) [P]. We did not obtain the original tender files. China’s November 2025 arms-control white paper calls for human beings to “take the ultimate responsibility” and for “safe, reliable and controllable military applications of AI” [G]. It names no programme-level simulation accreditor; that is a disclosure limit, not evidence that none exists.

What no source shows. We found no public military accident in which primary evidence attributes a failure to a gap between a training world and reality.

CHAPTER 04

Does the world predict reality?

For narrow tasks, sometimes yes: several studies find simulated and physical results that order robot policies in nearly the same way. But each figure comes from a particular evaluation — its tasks, policies, metric and adaptations — and several cited comparisons were produced by developers of the world, the benchmark or the policy; this issue has not quantified their share.

The strongest numbers, with their designs. In a 2024 study by UC San Diego, Stanford, Berkeley and Google DeepMind, SIMPLER's simulated scores for six policy checkpoints on the Google robot (three of them RT-1 checkpoints at different training stages), on three tasks, had an average task-wise Pearson correlation with real results of r = 0.924 [M]. PolaRiS, with authors from Physical Intelligence, Toyota Research Institute and six universities, reports a mean r = 0.90 across six paired scenes, 0.81 in the worst, after the policies were co-trained on simulated demonstrations [M]. REALM, from Czech and Dutch universities evaluating their own benchmark, reports r = 0.92 for task-progression scores in its Figure 6, supported by nearly 800 paired real and simulated rollouts [R].

Different evaluations, different numbers. In June 2026 a Tsinghua team evaluated three benchmarks on nine tasks and five newer policies, with 11,800 simulated and 1,115 real rollouts. Averaged over perturbation dimensions, SIMPLER's correlation was 0.402 (on seven of the nine tasks; two rotation tasks were not evaluated in SIMPLER), REALM's 0.785 and VLA-Arena's 0.725 [R]. The tasks, policies, perturbations and averaging differ from the original studies, so these figures cannot be read as one estimate of each simulator's fidelity, or as a before-and-after change. They show that a fidelity figure means little without the evaluation it came from.

— FIGURE 04 · SIM-VERSUS-REAL CORRELATION BY EVALUATION DESIGN

Reported correlations depend on the evaluation design

Pearson r between simulated and real results, one row per reported evaluation · not a league table, not a trend



Kaedim: error reduction 10.56 percentage points, 95% bootstrap CI 5.03–17.11; the change in r has an interval that includes zero.

CORRELATION IS ASSOCIATION, NOT CALIBRATION · AGIBOT'S R^2 NOT SHOWN (DIFFERENT STATISTIC) · SOURCE: CHAPTER 4

Ordering is not calibration. Kaedim, a company that sells 3D reconstruction, commissioned an outside lab to run and grade 200 physical trials (five tasks, two policies, ten task–policy means). Careful reconstruction cut the average gap between simulated and real normalised scores from 17.5 to 7.0 points: an improvement of 10.6 points with a 95% interval of 5.0 to 17.1. The correlation rose from 0.51 to 0.90, but the interval for that rise, −0.08 to 0.87, includes zero [M][V]. In the Stanford–Tsinghua Ctrl-World study, one policy closed a laptop in 70% of real trials and 5% of the world model's (the paper's Appendix B, Table 3); the authors note the model "tends to underestimate the execution success rate" [R].

Where it breaks. A Munich study of small physical cars found failures that simulation missed in a selected test: in one obstacle scenario, no failures in eight simulated runs and four in eight physical ones; perception was the main source of the gap (TUM/fortiss, 2025) [R]. Makers report similar limits for their own worlds: 1X lists contact violations, missed depth and false generated successes [M].

WHAT A STUDY MEASURES	WHAT IT CAN SUPPORT	WHAT IT CANNOT
Correlation between simulated and real scores for the same policies	Choosing between candidate policies on similar tasks	A probability that a chosen policy is safe
Absolute error in success rate or score	How far a simulated number can be trusted as a number	Transfer to new tasks or settings
Physical success after training in simulation	That the trained policy works on these tasks	That simulation, rather than other inputs, caused it, without a comparator
Agreement with an external physical leader - board	Quantitative plausibility across policies	Same-scene, same-policy fidelity

Who measured. Many comparisons in our inventory were produced by the developers of the world, the benchmark or the policy, including academic authors evaluating their own benchmarks. The Tsinghua and Munich studies are external evaluations of others' benchmarks or systems. We have not coded every entry for commercial affiliation, so we do not quantify the share. The assembled evidence does not meet the breadth our Lock sets for showing reliable prediction across a broad class — three independent teams, three kinds of robot, twenty predeclared task families — which leaves the question open rather than answered in the negative.

CHAPTER 05

The precedent: when a simulation counts as evidence

Self-driving, maritime autonomy, medical devices and nuclear power show what a public witness can look like: a bounded claim, a validated tool, a named assessor and a decision on record. In self-driving the rules are detailed; in the medical and nuclear comparisons we examined, the decision is published with the tool or method and its version or revision.

INSTRUMENT	WHAT IT REQUIRES OF A SIMULATION	WHO ACCEPTS	PUBLIC RECORD
Germany, AFGBV Annex 1, Part 2, No. 11	Simulation tools "must be validated" by comparison with "a representative selection of real trials"; "no significant difference" between simulated and real values; technical service may require real tests	Federal Motor Transport Authority (KBA) and technical services [G]	Approvals not published with toolchain detail in the records we found. "No significant difference" sets no numerical margin
EU, Regulation 2022/1426, fully automated vehicles	Annex III, Part 4: credibility of the virtual toolchain — fitness for purpose, configuration and version control, data pedigree, verification, validation, uncertainty; toolchain versions and validation thresholds preserved	Type-approval authority "shall assess the documentation" and may test physically [G]	Formal documentation separated from confidential technical material; no deployment's toolchain file retrieved
UN Regulation 157, automated lane keeping	Manufacturer establishes tool scope, scenario validity and correlation with physical tests	Technical service; type-approval authority [G]	Mercedes DRIVE PILOT approval by KBA with TÜV Rheinland as technical service is public; simulator version not retrieved
IMO MASS Code, autonomous ships; non-mandatory, governments invited to give effect from 1 Jul 2026	Simulation results "reviewed and accepted" where applicable, under an approved verification and validation plan	Flag Administration [G]	No certificate naming a simulator revision retrieved
ISO/PAS 34506:2026, published 7 Oct 2026	Qualifies virtual test environments for automated driving; validation of individual simulation models is outside this specification's scope	Voluntary unless referenced [G]	—
US federal level	No general type approval: manufacturers self-certify to safety standards; exemptions and state permits apply; NHTSA withdrew its proposed voluntary AV STEP programme on 26 June 2026 [G]	Companies; state regulators	California's CPUC approved Waymo's updated passenger safety plan and expanded service on 14 August 2026; the approval letter names no simulator or simulator version
US FDA, Medical Device Development Tools	Qualifies a computational tool for a stated context of use, so sponsors need not re-establish it	FDA [G]	"Virtual MRI Safety Evaluations of Medical Devices", version 1.0, qualified 16 Nov 2021, with its bench-test context of use and limits (summary); tool qualification is not device approval

INSTRUMENT	WHAT IT REQUIRES OF A SIMULATION	WHO ACCEPTS	PUBLIC RECORD
US NRC, reactor licence amendments	Plant-specific use of approved analysis methods	NRC [G]	Millstone Unit 3, Amendment 281 (7 Jan 2022): a fuel-centreline-temperature safety limit based on the fuel-melt values of Westinghouse's NRC-approved PAD5 model, Topical Report WCAP-17642-P-A Revision 1 (Federal Register, 22 Feb 2022, p. 9655); the safety evaluation limits PAD5's applicability and states that approving PAD5 for use at Millstone 3 was outside its scope

Company safety cases. Waymo publishes a safety-case method combining simulation, road data and engineering analysis, and reports that TÜV SÜD reviewed its safety-case management and remote-assistance programmes [M]; that is a process review, not a validation of the Genie 3-derived world. Aurora's safety case assigns release decisions to an internal Safety Review Board with the chief executive as accountable executive. Aurora commissioned Edge Case Research for a three-month review of its whole safety-case structure using a sampling approach that the assessor says it selected independently; Edge Case discloses that Aurora's chief executive sits on its board and states that board members do not review or influence report content [V][M]. Applied Intuition reports TÜV SÜD certification of its automotive virtual-validation platform (19 August 2026) [M].

What the precedent teaches. In the instruments reviewed, where regulators wanted simulation evidence they wrote down who accepts it and how the tool is checked against reality; EU type approval even requires toolchain versions and validation thresholds to be preserved. In the FDA and NRC comparisons the decision itself is published: a tool or method, a version or revision, a context of use, limits and an acceptor. In the self-driving cases we examined, we did not retrieve such a record for the simulator or world model; the military accreditation in chapter 7 has one at summary level, assembled from two reports.

CHAPTER 06

Rules and their exclusions

For civilian machinery in the EU, a notified-body route applies from 20 January 2027 to specified safety components with self-evolving machine-learning behaviour, and further AI requirements must apply by August 2028. The enacted text we reviewed does not establish public acceptance of a training world. The EU instruments use different military-related exclusions.

Civilian robots in the EU. The Machinery Regulation 2023/1230 applies from 20 January 2027. Annex I, Part A names "safety components with fully or partially self-evolving behaviour using machine learning approaches ensuring safety functions" and machinery with such embedded systems not placed on the market independently, in respect of those systems only; these go through a notified body. It also sets requirements on learning-related hazards, bounded behaviour and record-keeping; the manufacturer remains responsible for product safety [G]. A notified body may release the EU type-examination evaluation report "only with the agreement of the manufacturer", without prejudice to its duties to notifying authorities (Annex VII, point 5) [G]. The Regulation excludes weapons and machinery "specially designed and constructed for military or police purposes" (Art. 2(2)(d) and (l)). The Digital Omnibus on AI (Regulation 2026/1744, in force 27 July 2026) places machinery in the AI Act's sectoral route: further AI-specific requirements are to be set by delegated acts under the Machinery Regulation and must apply by 2 August 2028 [G]. The AI Act excludes systems used "exclusively for military, defence or national security purposes" (Art. 2(3)) — a use-based test, unlike the Machinery Regulation's design-based one.

Robot standards elsewhere. The standard for dynamically stable industrial mobile robots, including legged ones, ISO 25785-1, is a committee draft whose scope excludes consumer, medical and military robots; a revision of the service-robot standard ISO 13482 is at the final-draft stage [G]. South Korea's statutory certification route for outdoor mobile robots took effect on 17 November 2023 [G]. JQA reports that on 1 October 2026 it certified Mitsubishi Electric Mobility's hospital transport robot under the JIS B8445 mark scheme, the first such certification in Japan according to JQA's own database check as of 2 October (JQA) [V]. In the US, OSHA states that "there are currently no specific OSHA standards for the robotics industry", while general duties apply [G]. The scope, status and certification records reviewed do not establish a training-world acceptance route; we have not completed a clause-level review of all these standards.

China. A standards system for humanoid robots and embodied intelligence (2026 edition) was released on 28 February 2026 at a meeting organised by MIIT's humanoid and embodied-intelligence standardisation committee, with MIIT officials among the presenters; reports describe six areas including safety. We read only a government-hosted press account, not the framework itself [P]. A separate recommended national standard, GB/T 48316-2026, *Humanoid robots — Technical specification of simulation testing platform*, was issued on 28 September 2026 and applies from 1 April 2027 [G]. A State Administration for Market Regulation notice of 28 October 2026 lists it among standards whose full texts are made public [G]; we did not read its clauses and do not infer whether it names an acceptor. For automated driving, GB/T 47025-2026 sets recommended simulation-test methods (current since 28 January 2026) and the mandatory GB 44721-2026 applies from 1 July 2027 [G].

Military systems. In the US, DoD Directive 3000.09 requires autonomous weapon systems to undergo "rigorous hardware and software V&V and realistic system developmental and operational T&E", uses modelling and simulation for automated testing "whenever feasible", and requires senior review before formal development and again before field-

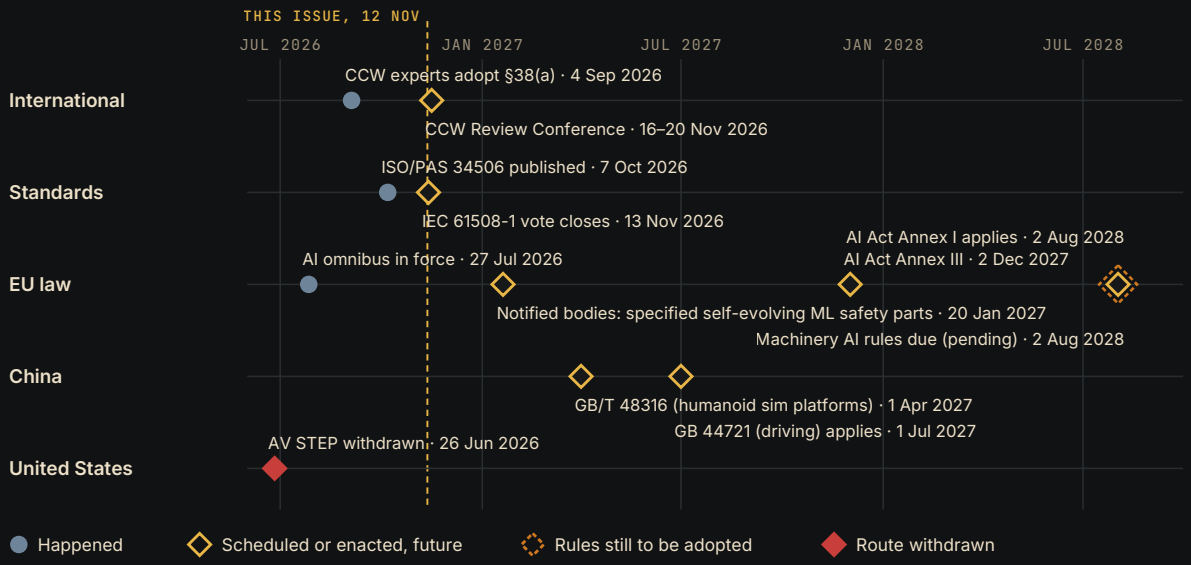
ing [G]. It covers weapon systems, with exceptions and an urgent-need waiver. Simulation accreditation sits in DoDI 5000.61 and DoDM 5000.102 (chapter 2). The UK assigns AI ethical assurance to named Responsible AI Senior Officers (JSP 936); Japan's defence acquisition agency's guideline provides for review of higher-risk AI projects at design and test milestones, with lower-risk projects largely self-assessed (ATLA guideline, June 2025) [G]. These are different objects — simulation accreditation, ethical assurance, risk-tiered review — and are not equivalent.

International. The CCW expert group's elements (adopted 4 September) are recommendations to the Review Conference; they do not displace existing international law, which continues to apply. UN General Assembly resolutions on lethal autonomous weapons passed in 2023, 2024 and 2025. On Resolution 80/57, adopted 1 December 2025 by 164 votes to 6 with 7 abstentions [G], the United States, Russia and Israel voted against, China and Saudi Arabia abstained, the UAE and Qatar voted in favour, and Ukraine recorded no vote (compiled recorded vote) [P]. On 29 October 2026 the General Assembly's First Committee approved draft resolution A/C.1/81/L.37 on lethal autonomous weapons systems by 163 votes to 5 with 9 abstentions, according to the UN's automatic meeting transcript, which is not an official record [G]; this is committee approval, not plenary adoption. The draft calls on CCW parties to advance to negotiations on an instrument and contains no provision on testing or simulation (draft text partner-checked; New Zealand's explanation of vote, 30 October, gives the aim as to "advance to negotiations on an instrument" [G]). The Review Conference's draft final declaration, posted on 28 October, leaves its paragraph and decision on lethal autonomous weapons as placeholders [G].

— FIGURE 05 · RULES AND DECISION POINTS, 2026–2028

EU machinery: ML safety parts from 2027, further AI rules by 2028

Dates of adoption, entry into force or application, June 2026 to September 2028 · status as of 23 Oct 2026



NO PUBLIC REGISTER OF WORLD ACCEPTANCE IN THE ENACTED TEXTS REVIEWED · SOURCE: CHAPTER 6

What applies when (2026–2028).

DATE	EVENT	STATUS AT THE CUTOFF (11 NOVEMBER 2026)
4 Sep 2026	CCW expert group adopts final report with “adequately realistic simulations” (§38(a))	Adopted; recommendations
7 Oct 2026	ISO/PAS 34506 published	Voluntary specification
29 Oct 2026	UN General Assembly First Committee approves draft A/C.1/81/L.37 on lethal autonomous weapons (163–5–9, automatic transcript)	Committee approval; plenary pending
13 Nov 2026	IEC 61508-1 final-draft vote closes	Scheduled, after this issue
16–20 Nov 2026	CCW Seventh Review Conference, Geneva	Scheduled; draft final declaration (28 Oct) leaves the LAWS text as placeholders
20 Jan 2027	Machinery Regulation applies; notified-body route for specified self-evolving ML safety components	Enacted, future application
1 Apr 2027	China GB/T 48316-2026, humanoid simulation testing platforms	Issued, recommended
1 Jul 2027	China GB 44721-2026, automated-driving safety	Issued, mandatory, not yet applicable
2 Dec 2027	AI Act high-risk obligations for Annex III systems	Enacted, future application
2 Aug 2028	AI Act Annex I obligations	Enacted, future application
by 2 Aug 2028	Delegated AI requirements for machinery must apply	No published draft or adopted act located (records searched to 11 Nov 2026)

CHAPTER 07

The mandate

The question is not whether anyone accepts a world, but whether an outsider can trace what was accepted. For each identified case, the register below records which links public sources supply and which they leave unresolved. One autonomy-adjacent case, the F-35 JSE accreditation, and three comparison cases — FDA, NRC and a US Army training-simulator accreditation — provide a complete link at summary level; we did not inspect their full dossiers.

Shadow 05 described the decision without a decider; Shadow 07 the change without a signatory. In Shadow 08 the signatory exists. For the self-driving cases, the public sources recovered leave the signature's content unresolved: which world version, which evidence, which limits, accepted by whom. When a system trained or tested in a world fails, an investigator will ask what was accepted, whether the conditions were exceeded, and who managed the boundary. The register shows how far public sources answer the first question today.

— FIGURE 06 · WHO HOLDS EACH STEP WHEN A SIMULATION IS USED AS EVIDENCE

Roles are named in the rules; public records vary by case

Who holds each step when a simulation is used as evidence · as read from the rules and records in chapters 2, 5, 6 and 7

	Self-driving EU, Germany	Civilian robots EU, from 2027	US military DoD	Medical tools FDA, comparison
Produce the evidence	Manufacturer, with tool vendors	Manufacturer	Programme with vendors; vendor may own the simulator	Tool developer
Assess the evidence	Technical service	Notified body, for specified ML safety parts	Operational or live-fire test agency; DOT&E	FDA reviewers
Accept for a use	Type-approval authority	Manufacturer declares; notified body certifies	Accreditation authority, named in the record	FDA qualification for a context of use
Public trace of what was accepted	Approval and approver public; simulator version not retrieved	Enacted text requires no public world record	Decision, use, limits and version: two reports (F-35 JSE)	Tool, version, use and limits published (one tool)

Named in a rule, or complete public record · Partial public record · Not located, or not required

PURPOSIVE SEARCH, NOT A SAMPLE · A ROLE IN A RULE IS NOT PROOF OF COMPLIANCE

The case register (fixed 23 October 2026; classes and version rule made explicit after round 3; C1 recoded after the pre-cutoff sweep). The register separates acceptance records involving simulation evidence (C1–C3), a candidate case whose eligibility is unresolved (C4) and contextual announcements (C5, C6). Only eligible acceptance records are tested. A record is *complete at summary level* if it links an identified simulation version, its intended use, a substantive evidence or limitations summary, and the accepting body and decision. A version is an explicit release or

build, an evaluated-method revision, or a named and dated frozen configuration baseline; a decision date alone does not count. This rule was made explicit after the results were known (Appendix C). Cell outcomes: *public*, *not located* or *not retrieved* (searched, not found), *not established*, *N/A*. Full search notes are in Appendix A.

ID	CASE AND DECISION	SIMULATION USED AS EVIDENCE?	VERSION	INTENDED USE	EVIDENCE OR LIMITS	ACCEPTOR
C1 · complete (summary)	F-35 Joint Simulation Environment, accredited 1 Sep 2023 (DOT&E FY2023, p. 43; FY2025, p. 33; AFOTEC)	Public	Public: F-35 cockpit configuration 30S02, "same as that used for IOT&E in FY23" (FY2025 report); an identifier for the aircraft model inside JSE, not a whole-environment manifest	Public: IOT&E	Public: "with limitations"; 65 deficiencies against baseline JSE requirements; limitations not listed	Public: F-35 Executive Committee
C2 · partial	Waymo service approval by CPUC, 14 Aug 2026; World Model announced 6 Feb 2026	Public in filings (testing and simulation discussed; partner-checked)	Not located	Public: passenger service in stated areas	Not located for any world	Public for the service (CPUC), not for a world
C3 · partial	Mercedes DRIVE PILOT, extended type approval up to 95 km/h (KBA; TÜV Rheinland technical service), Dec 2024	Public in maker material (virtual testing described; partner-checked)	Not located	Public: function and scope	Almost 150-page report described; full report and accepted simulator configuration not retrieved in this review	Public: KBA
C4 · eligibility unresolved; not assessed	Level 4 vehicle approval at Eiheiji, Japan, 30 Mar 2023	Not established	N/A	Public: vehicle, route, conditions	N/A	Public: Chubu District Transport Bureau
C5 · context; N/A	Shield AI acquisition of Aechelon, completed 22 Jun 2026	No acceptance decision (acquisition)	N/A	N/A	N/A	N/A
C6 · context; N/A	Applied Intuition selected for Autonomy Factory platform, 14 May 2026	No acceptance decision (platform selection)	N/A	Stated platform purpose: develop, validate, certify	N/A	N/A
X1 · complete (summary)	*Comparison:* FDA, Virtual MRI Safety Evaluations of Medical Devices, qualified 16 Nov 2021	Public	Public: version 1.0	Public: RF heating near passive implants, bench-test context, 1.5 T/3 T	Public: summary of evidence and limits	Public: FDA
X2 · complete (summary)	*Comparison:* NRC, Millstone 3 Amendment 281, issued 7 Jan 2022	Public (approved analysis method as the basis for a safety limit)	Public: WCAP-17642-P-A Rev. 1 (PAD5)	Public: fuel-centreline-temperature safety limit	Public: safety evaluation with applicability limits; plant-specific use of PAD5 outside its scope; full dossier not inspected	Public: NRC

ID	CASE AND DECISION	SIMULATION USED AS EVIDENCE?	VERSION	INTENDED USE	EVIDENCE OR LIMITS	ACCEPTOR
X3 · complete (summary)	*Comparison:* US Army AVCATT (Aviation Combined Arms Tactical Trainer), accreditation of software version 17 conducted September 2017 (Ferrell, *Army Aviation*, 31 Jul 2018, pp. 34–35)	Public (the simulator itself assessed, for human training)	Public: version 17; "Software version 17.0 partially accredited" in the Army's 2018 Weapon Systems Handbook, p. 134	Public: units' "collective training objectives"	Public: assessed "against our training doctrine" across plan, prepare, execute and assess; "not a flight task trainer"; accreditation partial	Public: Directorate of Simulation, US Army Aviation Center of Excellence

This is a purposive register, not a sample: it cannot estimate how often acceptance is public. The military record (C1) is complete at summary level, but only when two DOT&E reports are read together. The FY2023 report records the decision, the use, the acceptor and that limitations were attached; the FY2025 report, published in 2026, describing a planned second JSE site with "up to 12 F-35 cockpits in the 30S02 configuration (same as that used for IOT&E in FY23)", identifies 30S02 as the configuration used for IOT&E in FY2023 [G]. We coded C1 as partial until our pre-cutoff sweep recovered that sentence; the coding changed under the correction rule set in advance (Appendix C). Two cautions apply. The identifier names the aircraft model's configuration inside JSE, not a manifest of the whole environment, and JSE tests a human-piloted aircraft, not learned autonomy. Under a stricter rule requiring a whole-environment identifier, C1 would remain partial; we report both readings. DOT&E also names the configuration for the next accreditation, the 30S08 F-35-in-a-Box software for the 30R08 release, with formal accreditation for operational test planned for early FY2027, that is from October 2026, and at both JSE sites in the fourth quarter of FY2027 (FY2024, p. 48; FY2025, p. 33) [G]. That decision falls outside the fixed register; no record of it was located by the cutoff. The self-driving approvals we examined make the decision and the acceptor public but not, in the records we retrieved, the accepted simulation. One military case and two self-driving cases do not show which sector discloses more. The C1 record shows that a summary-level chain can be assembled from annual public reporting; the FDA, NRC and Army comparison records show the same for their sectors. AVCATT trains people, not autonomous systems, so it does not bear on H2.

Evidence production and acceptance. Applied Intuition was selected to build a platform intended for programme offices, developers and operators to develop, validate and certify autonomous systems, and reports a TÜV SÜD certificate for its automotive validation platform [M]. Shield AI now owns Aechelon's simulation business [M]. AgiBot evaluates its policies in its own simulator; NVIDIA co-authors evaluations in its own simulation stack [M]. DoDI 5000.61 assigns the acceptance decision to the authorised user or accreditation authority; supplying the tool does not by itself confer that role. Overlap between the builder and the evidence producer does not imply that any assessment was compromised. Where the public trace is incomplete, an outsider cannot tell from public records where evidence production ended and acceptance began.

Three ways this could go by 2028.

- *Public summaries spread.* Regulators borrow the FDA and NRC model, and the practice visible in DOT&E reporting and Army simulator accreditation: a short public record of tool, version, intended use, limits and acceptor, with confidential detail held by the authority. The EU's machinery delegated acts are one opportunity.
- *Platforms become the reference.* A few certified simulation platforms become the shared evidence base for many users, and the platform's certificate stands in for the world's acceptance.
- *Evidence after the fact.* Acceptance stays private until an incident forces disclosure, as crash investigations have in self-driving.

What would settle it. A minimal public acceptance summary for each world used as safety evidence: the world or toolchain version, the intended use and limits, a summary of validation evidence, and the accepting authority and date. The FDA, NRC and AVCATT comparisons, and the F-35 JSE record read across two reports, show it can be done without publishing source code. With such summaries, H1d could be tested on a register of deployments rather than a purposive set; measuring whether verification keeps pace with building would additionally need capacity and workload data.

HORIZON · SHADOW 09 PREVIEW

Next issue

> *Two candidates for Shadow 09 follow from this issue. One is the data under the world: combat footage, gameplay clips and teleoperation logs now shape what machines learn, and their provenance raises the same tracing question one layer down. The other is the insurer as witness: where acceptance records stay private, the parties that price the risk may ask for them.*

REGISTER

Appendix A. Case register: search notes

Records searched 20–23 October 2026 by the Foresight Future Institute and its research partner; searches repeated on 7 and 9–11 November 2026 without change to any coding. Outcomes apply to the records listed; they do not establish that a record does not exist.

ID	RECORDS FOUND	MISSING LINK AND SEARCH NOTE
C1	DOT&E FY2023 F-35 report, p. 43 ("The F-35 Executive Committee (EXCOM) accredited JSE, with limitations, for IOT&E on September 1st, 2023"); DOT&E FY2025, p. 33 ("up to 12 F-35 cockpits in the 30S02 configuration (same as that used for IOT&E in FY23)"), recovered by the research partner and opened by us on 23 Oct 2026; AFOTEC article on the accreditation	Recoded from partial to complete at summary level on 23 Oct 2026 (Amendment A4). The identifier is retrospective and names the aircraft-model configuration, not the whole environment; the specific limitations are not listed. JSE is a networked simulation for operational test and training of a human-piloted aircraft, not a learned-autonomy world. DOT&E's FY2024 (p. 48) and FY2025 (p. 33) reports identify the next JSE configuration (30S08 F-35-in-a-Box, for 30R08) and plan its accreditation for operational test from early FY2027; FY2025 describes a planned initial 30S08 capability (1QFY26) with limited regression testing "but not verification and validation (V&V) needed to accredit for OT"
C2	CPUC approval letter for Advice Letters 4 and 4-A (14 Aug 2026); the advice letters, which discuss testing and simulation (partner-checked); Waymo World Model and August 2026 posts	No record links an accepted simulation release to its evidence; the approval letter names no simulator
C3	TÜV Rheinland release (19 Dec 2024): KBA extended type approval, almost 150-page report; maker safety material describing virtual testing (partner-checked)	Accepted simulator configuration and credibility file not located; full report not retrieved. The release names test-track and public-road tests and does not mention simulation
C4	MLIT approval summary: Level 4 vehicle approval, 30 Mar 2023 (partner-checked)	Use of simulation as acceptance evidence not established: eligibility unresolved, not assessed
C5	Shield AI completion release (22 Jun 2026)	Context: an acquisition, not an acceptance decision
C6	Applied Intuition release (14 May 2026)	Context: a platform selection, not an acceptance decision
X1	FDA MDDT list (qualification 16 Nov 2021) and qualification summary	Bench-test context (analogous to ASTM F2182); tool qualification is not device approval
X2	Federal Register notice, 22 Feb 2022, p. 9655; NRC safety evaluation (ML21326A099), opened 23 Oct 2026	Limitation and condition (a) restricts PAD5's applicability; the evaluation states that approving PAD5 for use at Millstone 3 was outside its scope and that it does not approve PAD5 for associated parameters; full dossier not inspected
X3	COL John M. Ferrell, "Simulator Accreditation – AVCATT Update", *Army Aviation* 67(7), 31 Jul 2018 (printed issue date), pp. 34–35 (publisher PDF, opened 23 Oct 2026; reader); US Army Weapon Systems Handbook 2018, printed p. 134: "Software version 17.0 partially accredited", programme status 4QFY17 (PDF; page extract supplied by the research partner and read by us on 23 Oct 2026)	A first-person account by the responsible director, not the signed accreditation report; exact September day not retrieved. Human-training simulator, so a comparison, not an autonomy case. Earlier excluded for access; access resolved in round 3
Also searched	UK self-driving pilot guidance (31 Mar 2026); Beijing, Shenzhen and Wuhan testing rules and permit reports; IMO MASS Code; EASA AI concept paper	Rules or guidance; no completed autonomy acceptance record with a linked simulation version found

REGISTER

Appendix B. Actor ledger (27 actors)

One claim per actor, coded 23 October 2026; levels are not cumulative and do not rank companies. Recoded after round 3: Odyssey and Fourier (L2-T reported), Wayve and UBTech (L1, metric named), Meta (physical planning, recorded separately). After the pre-cutoff sweep: Runway (L2-P), Decart (L1), AgiBot figures from v4. Company-coauthored papers count as the maker's record. Financing details are in chapters 1 and 2.

ACTOR	LEVEL (OUR CODING)	EVIDENCE ITEM AND DATE	BOUNDARY	SOURCE
World Labs	L2-P	2026-07-28 published. Bimanual cube handover; 2,000 simulated and 100 real trials per checkpoint; ranking chart, no scalar metric.	Limited tasks/checkpoints; not blanket Marble certification	World Labs
AMI Labs	NR	2026-10-23 checked. Released quantitative physical-transfer item not resolved.	Research agenda/funding cannot establish either successful or failed verification	AMI Labs
Emulate	NR	2026-10-23 checked. Released model or physical-transfer study not resolved.	A public demo was not established either; do not plot NR as L0	pymnts.com
General Intuition	L1	2026-10-23 checked. MIRA game-world performance claims.	Game performance is not a physical-transfer or world-fidelity result	General Intuition
Physical Intelligence	L2-P	2025-12-18 published. PolaRiS coauthored paired evaluation, D11.	Calibration/co-training matters; not validation of every pi model	Paper
Skild AI	L1	2026-08-18 published. S1 internal cumulative per-step success comparisons.	Per-step success is not end-to-end reliability or paired physical calibration	Skild AI
Figure	L1	2026-01-27 published. Helix 02 reported continuous demonstration duration/action count.	No success-rate denominator or simulation/physical predictive calibration	Figure
1X	L2-P	2025-06 published. World-model checkpoint ranking against blind real A/B runs; held-out success/failure prediction accuracy 63.06% → 71.17%.	Selected tasks; video human ratings are not robot trials; disclosed hallucination/error limits	1X · 1X
Apptronik	L0	2026-10-23 checked. Apollo public physical demonstrations/training collaborations.	Customer presence is not simulator validation; quantitative paired result NR	Apptronik · Apptronik
Generalist AI	L1	2026-04-02 published. GEN-1 selected real-task success metrics.	Human-data pretraining is not simulation; real-task success alone does not establish L2	Generalist
Rhoda AI	L1	2026-03-10 published. Industrial workflow time/no-intervention claim.	Customer, trial denominator and uncertainty NR; no paired fidelity result	Rhoda
FieldAI	L1	2025-08-20 published. Maker reports hundreds of industrial environments.	Deployment count is not success rate or simulator predictivity	FieldAI
Decart	L1	2026-06-10 published. Oasis 3 specification: "22 FPS at 512×768×3, with under 200 ms of latency".	Specifications are not fidelity or transfer results	Decart

ACTOR	LEVEL (OUR CODING)	EVIDENCE ITEM AND DATE	BOUNDARY	SOURCE
Odyssey	L2-T reported	2026-09-15 published. Policies trained entirely in simulation "traveled about 77% as far between safety-driver interventions as those trained on real footage".	Simulator identity and version, physical exposure and uncertainty NR; not a paired predictive comparison	Odyssey
Runway	L2-P	2026-02-27 published. GWM-Robotics: eight policies, 1,450 simulated rollouts against RoboArena physical results; Pearson r 0.95 for policy ranking, MMRV 0.033; D24.	Ranking, not calibration; physical trial count and interval NR; one arm, tabletop tasks	Runway
Wayve	L1	2026-08-03 published. GAIA-4: training improves how faithfully it preserves the recorded world "by 2.5x"; outcome comparisons shown qualitatively.	No absolute score, baseline or task-matched physical-outcome comparison	Wayve
Waabi	L2-P	2025-03-11 published. Waabi World paired trajectory-realism metric, D05.	Vendor-defined score is not safety/reliability; denominator and uncertainty NR	Waabi
Applied Intuition	L3 reported	2026-08-19 published. Maker reports TÜV SÜD VIVALDI general-readiness assessment.	Certificate record/number/grant date unresolved; not a named accepted safety case (L4 NR)	Applied Intuition
Google DeepMind — Genie 3	L1	2025-08-05 published. Genie 3 generated-world consistency/interaction claims.	No direct physical comparator resolved; judge Waymo derivative separately. Google's Argon MuJoCo transfer result (13 Oct 2026, T07) is a separate system	Google DeepMind
NVIDIA and co-authors (SimFoundry)	L2-P	2026-08-05 (v4). SimFoundry real-to-sim evaluation, 7 tasks, 5 policy architectures; mean Pearson r 0.911, MMRV 0.018.	Vendor-coauthored study; Cosmos simulator-to-simulator trials are not new physical trials	SimFoundry, arXiv 2606.28276
Meta — V-JEPA 2	Physical planning (separate)	2025-06-11 published. V-JEPA 2-AC, learned from real robot data, plans in two labs; success out of ten trials with varied initial conditions, T03.	Not training in a simulated world (recoded from L2-T); not matched fidelity	Paper
Tesla — Optimus and simulation	L0	2026-10-23 checked. Optimus public capability presentations.	FSD fleet metrics do not validate the Optimus training world; link now redirects, replacement footage not re-viewed (provisional)	Tesla
Unitree	L1	2026-10-23 checked. UnifoLM internal task/data-volume claims.	External research on Unitree hardware does not validate the company's entire model family	Unitree
UBTech	L1	2026-10-23 checked. Walker S2 product claim: "swap battery autonomously within 3 minutes".	A task-duration metric, not evidence about simulation	UBTech
AgiBot	L2-P/T	2026-08-14 (v4; first version 5 Jan 2026). Genie Sim 3.0: 16 cells, 50 real and 50 simulated trials each; R^2 0.931, slope about 1.02; D13.	Selected tasks/same base model; R^2 is not Pearson r; CI NR; v1 reported R^2 0.924	Paper
Galbot	L2-T	2025-05 published. GraspVLA synthetic-training transfer, T02.	Success permits three attempts; vendor/university coauthorship; not predictive calibration	Paper
Fourier	L2-T reported	2024-11-06 published. GR-2 floor-to-stand policy trained in Isaac Gym "achieved an 89% success rate" on physical controls.	NVIDIA-published account [M][V]; physical trial count and interval NR	NVIDIA

REGISTER

Appendix C. Interpretation Lock, summary

Version 1.0 (frozen 11 November 2026: Amendments A1–A4 of 23 October and logs of the 7 November sweep and the final check), revised after three adversarial rounds and two pre-cutoff sweeps; A3 holds clarifications made after the case results were known; A4 records the C1 recoding and its consequence for H1d. Not preregistered. This is a summary; the full Lock is published with this issue.

TEST	REJECTS	STATUS
H-1: ≥80% of a frozen eligible set of actors publish L2-P evidence for a material claimed task, with version, physical n, uncertainty, holdouts and failures	H1a	Not computed: eligibility and material-task rule not frozen; the 7 of 27 count in chapter 1 is not this test
H-2: an applicable instrument names an assessor or acceptor for robot learned behaviour	Former H1c robot clause	Met by the Machinery Regulation (enacted; applies from 20 Jan 2027)
H-3: ≥3 organisationally independent teams, ≥3 embodiment classes, ≥20 predeclared task families, held-out policies and shifts, lower 95% bound of correlation ≥0.70, plus calibration and failure-detection limits	A claim that reliable prediction is absent across a broad class (not made)	Not met by current evidence
H-4a: an eligible listed case has a public record, published by the cutoff, linking an identified version (chapter 7 rule), intended use, a substantive evidence or limitations summary, and the acceptor and decision; a qualifying record found later still counts	H1d for that case	Met at summary level for C1 (two DOT&E reports; aircraft-model identifier) and for comparisons X1–X3; not met for C2–C3; C4 not assessed; C5–C6 N/A. H1d is therefore rejected as stated and retained only for C2–C3
H-4b: ≥90% of ≥30 sampled deployments have such a record, public or attested	A broader prevalence claim (not made)	Not computed
H-5: matched capacity and workload data over ≥3 years	H1b	Not computed
H-6: civilian buyers >60% of paid demand for training and test environments after unknowns counted as military	H2 demand	Not computed
H-7: in ≥12 frozen matched capabilities, civilian first operational acceptance ≥12 months before military in ≥2/3, with civilian lineage	H2 timing	Not computed
H-8: in ≥12 matched rule pairs (same capability, object, jurisdiction, legal force), civilian binding rule first in ≥2/3	H2 rule origin	Not computed

REGISTER

Appendix D. Evidence inventory (dated)

Inventory assembled by the research partner and checked 23 October 2026; two entries added in round 3 and two in the pre-cutoff sweep. D = predictive or comparative simulated-versus-physical entries; T = further physical-outcome entries; B = benchmark diagnostics that are not matched physical policy validation. Entries are not independent studies: one paper can supply several, and many are authored by the developers of the world or benchmark. Recoding log: T03 (Meta V-JEPA 2-AC) recoded from training transfer to physical planning; T05 and T06 added; D24 (Runway GWM-Robotics) and T07 (Google Argon) added and D13 updated to v4 in the sweep; B01 (Physics-IQ, filmed physical experiments) and B04 (NVIDIA Cosmos, simulator-to-simulator) kept as diagnostics, outside the L2-P count. Total: 24 D + 7 T + 5 B = 36 entries.

ID	LABEL	DATE, ITEM AND SUBTYPE	SOURCE
D01	[R]	11 Apr 2023 published, Learning naturalistic driving environment with statistical realism / distribution fidelity	link
D02	[R]	2 May 2023 journal publication, Model vs System Level Testing of Autonomous Driving Systems: A Replication and Extension Study / comparative testing	link
D03	[M]	3 Oct 2023 preprint; 13 Jul 2024 v2, What Do We Learn from a Large-Scale Study of Pre-Trained Visual Representations in Sim and Real Environments? / L2-P method ranking, plus L2-T	link
D04	[M]	9 May 2024 preprint, Evaluating Real-World Robot Manipulation Policies in Simulation (SIMPLER) / L2-P	link
D05	[M]	11 Mar 2025 maker research post, Waabi World realism / L2-P published maker metric	link
D06	[M]	4 Apr 2025 preprint, Real-is-Sim / L2-P	link
D07	[M]	June 2025 paper/release, 1X World Model: Evaluating Bits, not Atoms / L2-P	link
D08	[R]	26 Sep 2025 preprint, A Multi-Modality Evaluation of the Reality Gap in Autonomous Driving Systems / L2-P	link
D09	[R]	11 Oct 2025 preprint; 1 Mar 2026 v3 / ICLR 2026, Ctrl-World / L2-P and L2-T	link
D10	[M]	6 Nov 2025 preprint; 10 Nov v2, Real-to-Sim Robot Policy Evaluation with Gaussian Splatting Simulation of Soft-Body Interactions / L2-P	link
D11	[M]	18 Dec 2025 preprint; 30 Dec v2, PolaRiS / L2-P	link
D12	[R]	22 Dec 2025 preprint, REALM / L2-P	link
D13	[M]	5 Jan 2026 preprint; 14 Aug 2026 v4, Genie Sim 3.0 / L2-P and L2-T	link
D14	[M]	11 Feb 2026 preprint, MolmoSpaces / L2-P aggregate external benchmark	link
D15	[M]	9 Mar 2026 preprint, Interactive World Simulator / L2-P and L2-T	link
D16	[M]	Apr 2026 preprint; 14 Aug v4, RoboLab / L2-P aggregate external benchmark	link
D17	[R]	9 Jun 2026 preprint, A Practical Recipe Towards Improving Sim-and-Real Correlation for VLA Evaluation / L2-P, cross-benchmark assessment	link
D18	[M]	26 Jun 2026 preprint/project, SimFoundry / L2-P	link

ID	LABEL	DATE, ITEM AND SUBTYPE	SOURCE
D19	[M]	1 Jul 2026 preprint, RoboWorld / L2-P learned-world aggregate benchmark	link
D20	[M]	28 Jul 2026 maker research publication, World Labs Building Worlds That Train Robots / L2-P and L2-T	link
D21	[M]	21 Sep 2026 preprint; 22 Sep v2, H2RBench / L2-P method ranking	link
D22	[M] [V]	30 Sep 2026 submitted; 2 Oct v2, Kaedim Measuring Asset and Scene Reconstruction Effects in Real-to-Sim Robot Evaluation / L2-P and scoped third-party test	link
D23	[R]	2025 journal publication, Olleja/Markkula/Bärgman, Validation of human benchmark models for automated driving system approval: How competent and careful are they really? / component-model fidelity	link
D24	[M]	27 Feb 2026 maker research post, Runway GWM-Robotics policy evaluation against RoboArena / L2-P ranking; added in the sweep	link
T01	[M]	30 Aug 2023 published, Champion-level drone racing using deep reinforcement learning (Swift)	link
T02	[M]	6 May 2025 preprint, GraspVLA	link
T03	[M]	11 Jun 2025 preprint, V-JEPA2 / V-JEPA2-AC real planning	link
T04	[M]	12 Jan 2026 maker experiment publication, 1X World Model Self-Learning	link
T05	[M] [V]	6 Nov 2024 published, Fourier GR-2 floor-to-stand policy trained in Isaac Gym (NVIDIA account); added in round 3	link
T06	[M]	15 Sep 2026 published, Odyssey-3: driving policies trained in simulation, physical intervention distance; added in round 3	link
T07	[M]	13 Oct 2026 maker post, Google Argon: physical success after MuJoCo simulation training; added in the sweep	link
B01	[M]	14 Jan 2025 Physics-IQ preprint; 17 Jun 2026 Physics-IQ-Verified preprint	link
B02	[M]	27 Feb 2025 released; NeurIPS 2025, WorldModelBench	link
B03	[R]	27 Oct 2025 preprint; ICLR 2026 published, RobotArena∞	link
B04	[M]	21 Oct 2026 technical report released, NVIDIA Interactive World Simulation with Cosmos for Physical AI	link
B05	[M]	19 Oct 2026 preprint, RoboValue	link

REGISTER

Sources

Primary records were opened between 20 and 23 October 2026 unless marked. Items checked by the research partner and not re-opened by us, failed retrievals and claim-by-claim sources are listed in the evidence appendix.

RULES AND OFFICIAL RECORDS

01 CCW Group of Governmental Experts on LAWS, final report CCW/GGE.1/2026/3 (adopted 4 Sep 2026)

02 DoD Instruction 5000.61, VV&A of models and simulations (17 Sep 2024)

03 DoD Manual 5000.102, M&S VV&A for OT&E and LFT&E (9 Dec 2024)

04 DoD Directive 3000.09, Autonomy in Weapon Systems (25 Jan 2023)

05 Germany, AFGBV Annex 1 · EU Implementing Regulation 2022/1426 · IMO MASS Code (MSC 111/22, Annex 16)

06 EU Machinery Regulation 2023/1230 · EU Regulation 2026/1744, Digital Omnibus on AI

07 ISO/PAS 34506:2026 · ISO/CD 25785-1 · OSHA, robotics standards

08 NHTSA, AV STEP withdrawal, 91 FR 38619 (26 Jun 2026)

09 FDA, Medical Device Development Tools · FDA, Virtual MRI Safety Evaluations, qualification summary · NRC notice, Federal Register, 22 Feb 2022

10 FAA, flight simulation device qualification (25 Sep 2002) · NRC, LOCA requirements history (7 Nov 2005)

11 UK JSP 936 Part 1 · Japan ATLA responsible AI guideline · China arms-control white paper (27 Nov 2025)

12 Ukraine Ministry of Defence: Bravel Dataroom · battlefield data (12 Mar 2026) · Avengers Labs · UK: partnership · competition · Q&A

13 US Army, xTechHumanoid results (30 Sep 2026)

14 Gulf: WAM on Tawazun · Gulf News on Tawazun · QNA · SPA · QIA on Applied Intuition · QIA on Apptronik

15 NRC, Millstone 3 Amendment 281 safety evaluation (ML21326A099) · Ferrell, “Simulator Accreditation – AVCATT Update”, *Army Aviation*, 31 Jul 2018 · US Army Weapon Systems Handbook 2018, p. 134

16 DOT&E, FY2023 Annual Report, F-35 (p. 43) · FY2024 (p. 48) · FY2025 (p. 33) · AFOTEC, JSE accreditation

17 China: standards system press account (23 Mar 2026) · GB/T 48316-2026

STUDIES

18 NVIDIA and co-authors, SimFoundry (v4, Aug 2026)

19 SIMPLER (May 2024) · PolaRiS (Dec 2025) · REALM (Dec 2025) · A Practical Recipe Towards Improving Sim-and-Real Correlation (Jun 2026)

20 Kaedim, real-to-sim reconstruction effects (Sep–Oct 2026) · AgiBot Genie Sim 3.0 (v4, Aug 2026) · Ctrl-World (Oct 2025) · TUM/fortiss reality gap (Sep 2025) · NVIDIA SimFoundry

21 Odyssey-3 (15 Sep 2026) · NVIDIA on Fourier GR-2 (6 Nov 2024) · Wayve GAIA-4 (3 Aug 2026) · Runway GWM-Robotics (27 Feb 2026) · Google Argon (13 Oct 2026) · Decart Oasis 3 (10 Jun 2026)

22 World Labs, Building Worlds That Train Robots (28 Jul 2026) · 1X World Model · Waabi World realism

23 CSIS, Russia’s drone ecosystem (13 Apr 2026)

COMPANIES AND PRESS

24 Waymo: World Model (6 Feb 2026) · 10 AI lessons (26 Aug 2026) · independent audits (Nov 2025)

25 TÜV Rheinland on DRIVE PILOT (19 Dec 2024) · JQA, JIS B8445 certification (2 Oct 2026)

26 Shield AI completes Aechelon acquisition (22 Jun 2026) · Edge Case Research on Aurora (Jun 2026)

27 Applied Intuition: Autonomy Factory (14 May 2026) · DoD BPA (21 Aug 2024) · TÜV SÜD VIVALDI (19 Aug 2026)

28 Funding and deals: World Labs · AMD–World Labs (28 Sep 2026) · AMI Labs (Reuters) · Skild AI · Wayve extension (15 Apr 2026) · Odyssey · Anduril · Helsing · Saronic · Waabi · Bloomberg on Physical Intelligence (27 Mar 2026)

29 Breaking Defense on the IDF branch (10 Sep 2026) · Reuters, PLA and AI (27 Oct 2025) · Reuters via Asharq Al-Awsat (7 Sep 2026)

REGISTER

Glossary

Evidence labels: [M] maker record · [V] vendor, assessor or investor with an interest · [G] government or standards body · [R] research · [P] press or secondary · NR not resolved.

World (in this issue) The object of acceptance in this issue: a model or toolchain at a stated version — a physics simulator, a digital twin, a generative world model or a synthetic-sensor pipeline. Real-world data, including combat data, is an input used to build, train or calibrate such worlds; access to data is not itself a world.

Version (in the case register) For the completeness test, a version is an explicit release or build identifier, an evaluated-method revision, or an explicitly identified, named and dated frozen configuration baseline. A decision date alone does not establish a configuration. This rule was made explicit after the case results were known and is logged as a post-result clarification. The F-35 JSE case meets it through a configuration identifier for the aircraft model inside JSE (30S02, named in a later DOT&E report); a stricter rule requiring a whole-environment identifier would leave that case partial.

World model A learned model that predicts how an environment responds to an action, so a machine can try options before it moves or be tested in scenes that never happened. Examples in this issue: Genie 3, Cosmos, V-JEPA 2, Waymo World Model. A world model is a kind of world; not every simulator is learned.

Digital twin A digital representation of a specific physical system or place, kept in step with it. US defence test rules (DoDM 5000.102) list digital twins among models that must be accredited before their results count as test evidence [G].

Sim-to-real (and real-to-sim) Sim-to-real: whether what a system learns or scores in simulation carries over to the physical world. Real-to-sim: rebuilding a real scene in simulation so that policies can be evaluated there. Both are measured, not assumed; results depend on the tasks and policies tested.

Evidence ladder (L0–L4) L0 demonstration · L1 internal metric claimed · L2-T quantitative physical results after training in a simulated world · L2-P quantitative comparison of simulated and physical task outcomes or trajectories for the same policies · L3 independent replication or scoped third-party test · L4 accepted in a named safety case by a regulator, certifier or military test authority. NR (not resolved) is kept apart from L0. Physical planning with a world model learned from real data, and generated-video comparisons with filmed experiments, are recorded separately. The ladder rates a claim, not a company.

Correlation vs calibration A Pearson correlation between simulated and real scores says how closely they move together; rank measures say whether policies are ordered the same way. Calibration says whether a simulated success rate of 80% means about 80% in reality. A world can rank well and still be consistently optimistic or pessimistic. Neither is a probability that a system is safe.

V&V and accreditation Verification asks whether a model was built as specified; validation whether it represents reality well enough for its intended use. Accreditation, in DoDI 5000.61, is "the official certification that a model, simulation, or distributed simulation is acceptable for use" for a specific purpose, recorded with the version, the intended use and the accepting authority [G]. Some accreditation decisions are reported publicly: the Pentagon's test office reports that the F-35 Joint Simulation Environment was accredited, with limitations, for operational test on 1 September 2023 [G]; an Army article by the responsible director reports the September 2017 accreditation of the AVCATT training simulator, version 17, for collective training [G].

Intended use The specific decision, conditions and limits for which a model or simulation is accepted. This is the purpose of V&V, not a flaw: FDA guidance and ASME V&V 40 call it the context of use. What matters for outsiders is whether the use, evidence and limits are visible.

Safety case A structured argument, supported by evidence, that a system is acceptably safe for a given use. In self-driving, simulation results can be part of that evidence. An L4 rating in this issue requires an identified authority to accept specified world evidence in a named safety case.

Type approval and technical service The EU and UNECE route by which a national authority approves a vehicle type before sale, on evidence assessed by a designated technical service. For automated vehicles it can include the credibility of the simulation toolchain (EU 2022/1426, UN R157) [G].

Notified body An organisation designated by an EU member state to assess conformity within a defined scope. Under the Machinery Regulation, this notified-body route applies from 20 January 2027 to specified safety components with fully or partially self-evolving behaviour using machine learning that ensure safety functions, and to machinery with such embedded systems not placed on the market independently, in respect of those systems only. Further AI requirements must apply through delegated acts by 2 August 2028 [G].

CCW and LAWS The Convention on Certain Conventional Weapons (CCW) is the UN framework where states discuss lethal autonomous weapons systems (LAWS). Its Group of Governmental Experts adopted a final report on 4 September 2026; the Seventh Review Conference is scheduled for 16–20 November 2026 in Geneva. The elements are recommendations; states retain responsibility under existing international law [G].

Digital Omnibus on AI Regulation (EU) 2026/1744, published 24 July and in force 27 July 2026. It deferred AI Act high-risk obligations (Annex III to 2 December 2027; Annex I to 2 August 2028) and placed machinery in the AI Act's sectoral route: AI-specific requirements come through delegated acts under the Machinery Regulation, which must apply by 2 August 2028; we did not locate a published draft or adopted delegated act in the records searched through the cutoff (11 November 2026) [G].

UN Regulation 157 The UNECE regulation for automated lane-keeping systems. It requires manufacturers to establish the scope and validity of simulation tools and their correlation with physical tests, assessed by a technical service for type approval [G].

ISO/PAS 34506:2026 Published 7 October 2026: qualification of virtual test environments for automated driving systems. Its scope "does not apply to the validation or assessment of individual simulation models" nor to evaluating driving performance [G].

NR — not resolved A field we could not resolve from the sources reviewed, with the reason kept. It means the information was not found, not that the thing did not happen. Internal validation may exist where no publication does.

IQT (In-Q-Tel) An independent not-for-profit investor established by the CIA. It appears as "IQT" among investors in Skild AI's Series C (January 2026) and Odyssey's Series B (June 2026) [M]. An investor is not a buyer.

Not an allegation against any named body. Nothing here constitutes legal, investment or security advice. Statements of fact are limited to matters recorded in the public sources cited and are reports of what those sources state. Prepared independently and without external funding by the Foresight Future Institute. AI assistance and adversarial review are disclosed in How to read this issue.